

Siegfried Bergs

OPTIMALITÄT BEI CLUSTERANALYSEN

Experimente zur Bewertung
numerischer Klassifikationsverfahren

1981

1981



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41555410_0032

OPTIMALITÄT BEI CLUSTERANALYSEN

EXPERIMENTE ZUR BEWERTUNG
NUMERISCHER KLASSIFIKATIONSVERFAHREN

Inauguraldissertation
zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaft
durch den
Fachbereich Wirtschafts- und Sozialwissenschaften
der Westfälischen Wilhelms-Universität Münster

vorgelegt von
Siegfried Bergs

aus Hamm

1980



Erster Berichterstatter: Prof. Dr. J. Bleymüller
Zweiter Berichterstatter: Prof. Dr. H. Gülicher
Dekan: Prof. Dr. D. Ahlert
Tag der mündlichen Prüfung: 4. Februar 1981

INHALTSVERZEICHNIS

	Seite
Verzeichnis der Tabellen	IV
Verzeichnis der Abbildungen	VI
Verzeichnis der wichtigsten Symbole	IX
1 Einführung	1
1.1 Clusteranalyse als multivariates Analyse- verfahren	1
1.2 Ziele der Clusteranalyse	10
2 Methoden der Clusteranalyse	19
2.1 Allgemeines	19
2.2 Hierarchische Verfahren	27
2.3 Iterative Verfahren	34
3 Probleme der Clusteranalyse	51
3.1 Auswahl und Aufbereitung der Daten	51
3.2 Distanz- und Ähnlichkeitsmaße	63
3.3 Die Kombination von Verfahrensbestandteilen	74
4 Die Analyse von Clusterstrukturen	83
4.1 Simulationsexperimente zur Strukturanalyse	83
4.2 Heuristische Regeln zur Auswahl der Clusterlösung	92
4.3 Tests auf Signifikanz der Clusterlösung	107
4.4 Zusammenfassung	121
5 Die Klassifikation von Objektmengen	123
5.1 Gütemaße für Klassifikationsverfahren	123
5.2 Experimente zur Klassifikation	130
5.3 Zusammenfassung	154
6 Schlußbemerkung	159
Literaturverzeichnis	163

VERZEICHNIS DER TABELLEN

	Seite
Tab. 2.1: Parameter von 8 hierarchischen Verfahren nach der Formel von LANCE&WILLIAMS	32
Tab. 2.2: STIRLING'sche Zahlen zweiter Art	45
Tab. 2.3: BELL'sche Zahlen	45
Tab. 3.1: Anzahl der übereinstimmenden und nicht- übereinstimmenden Merkmalsausprägungen bei zwei Objekten mit dichotomen Merk- malen	68
Tab. 3.2: Eigenschaften hierarchischer Verfahren	76
Tab. 3.3: Eigenschaften von Zielfunktionen	77
Tab. 3.4: Transformationen zur Datenaufbereitung	78
Tab. 3.5: Distanz- und Ähnlichkeitsmaße für me- trisch skalierte Merkmale	79
Tab. 3.6: Ähnlichkeitsmaße für nichtmetrische Merkmalssätze	80
Tab. 4.1: Durch Elbows hervorgehobene Cluster- zahlen	94
Tab. 4.2: Anzahl richtig zugeordneter Elemente	95
Tab. 4.3: Abweichungen des Varianzkriteriums zwi- schen Suboptimum und Mittelwert bzw. 5%-Punkt der empirischen Verteilung der Zufallswerte bei Datensatz VI nach einer z-Transformation	98
Tab. 4.4: Clusterzahlen mit der maximalen Diffe- renz zwischen Mittelwert bzw. 5%-Punkt der empirischen Verteilung der Zufalls- werte und dem suboptimalen Wert	100
Tab. 4.5: Clusterzahlen mit der maximalen Diffe- renz zwischen dem Kriteriumswert beim untersuchten Datensatz und bei den un- strukturierten Vergleichsdatsätzen	104

	Seite
Tab. 4.6: Clusterzahlen mit dem größten Wert von $K_1(i, i+1)$ bzw. $K_1^*(1, j)$	112
Tab. 4.7: Parameter der empirischen Verteilung von $K_1(j)/K_1(1)$ bei strukturierten und homogenen Datensätzen sowie Werte der KOLMOGOROV - SMIRNOV-Prüfgröße D bei Anpassung einer Normalverteilung und der kritische Wert d_c	116
Tab. 4.8: Mittelwerte der Nullverteilungen	118
Tab. 4.9: Standardabweichungen der Nullverteilungen	119
Tab. 4.10: Clusterzahlen mit der stärksten Abweichung vom Mittelwert der Nullverteilung	120
Tab. 5.1: Größte Werte des TANIMOTO-Koeffizienten beim Vergleich der richtigen Partition mit den künstlichen Partitionen	153

VERZEICHNIS DER ABBILDUNGEN

	Seite
Abb. 1.1: Beispiel zur Wirkung von Clusteranalyse, Diskriminanzanalyse und Identifikation	7
Abb. 2.1: Beispiel mit monothetischer und polythetischer Aufteilung	23
Abb. 2.2: Darstellung eines Dendrogramms bei agglomerativer und divisiver Zerlegung und Partition mit 3 Clustern	24
Abb. 2.3: Programmablaufplan eines Austauschalgorithmus, der zur Optimierung von 8 verschiedenen Optimalitätskriterien geeignet ist. Der Teil zur Berechnung des Kriteriumswertes ist in Abb. 2.4 gesondert dargestellt.	48
Abb. 2.4: Teil des Programmablaufplans in Abb. 2.3 zur Berechnung des Kriteriumswertes	49
Abb. 3.1: Erkennung eines Ausreißers im Dendrogramm	52
Abb. 3.2: Unterschiedliche Clusterung von KFZ im Variablenraum und im Faktorenraum	57
Abb. 3.3: Unterschiedliche Clusterung von KFZ bei Messung eines Merkmals in verschiedenen feinen Maßeinheiten	58
Abb. 3.4: Punkte gleichen Abstands von x_0 bei Distanzmessung nach (3.13)	66
Abb. 3.5: Orientierungshilfe zur Kombination von methodischen Bausteinen bei Clusteranalysen	82
Abb. 4.1: Scatterdiagramme der verwendeten zweidimensionalen Zufallsdatensätze	89
Abb. 4.2: Datensatz III nach z-Transformation	91

	Seite
Abb. 4.3: Entwicklung des Varianzkriteriums in einer Objektmenge mit drei Clustern (●—●—●) und einer Objektmenge ohne Cluster (○—○—○)	93
Abb. 4.4: Suboptimale Werte des Varianzkriteriums sowie Mittelwert und 5%-Punkt der empirischen Verteilung der Zufallswerte bei Datensätzen mit 125 Objekten und 1 bzw. 5 Clustern	99
Abb. 4.5: Datensatz VII und Vergleichsdatsatz mit einem Cluster	101
Abb. 4.6: Varianzkriterium bei Datensatz VI (●—●—●) und einem Vergleichsdatsatz (○—○—○)	102
Abb. 4.7: Empirische Verteilung des Varianzkriteriums bei Zufallsclusterung (I) und optimaler Clusterung (II) unstrukturierter Datensätze und bei optimaler Clusterung (III) strukturierter Datensätze	109
Abb. 4.8: Untere 1%-Punkte der Nullverteilung des WILKS'schen Lambda (○—○—○) und Werte des WILKS'schen Lambda bei Datensatz II (●—●—●)	110
Abb. 5.1: Versuchsobjekte in Versuch 1	132
Abb. 5.2: Häufigkeitsverteilung der Gruppenzahlen in Versuch 1	133
Abb. 5.3: Tatsächliche und wahrgenommene Länge in Versuch 1	134
Abb. 5.4: Transformierte und wahrgenommene Länge in Versuch 1	135
Abb. 5.5: Häufigkeitsverteilung der Gruppenzahlen in Versuch 2	136
Abb. 5.6: Streuungsdiagramm von Winkeln und Kathetenlängen in Versuch 3	137
Abb. 5.7: Anordnung der Dreiecke im Versuch 3	138
Abb. 5.8: Häufigkeitsverteilung der Gruppenzahlen in Versuch 3	139

	Seite
Abb. 5.9: Häufigkeiten h_{ij} , mit der die Dreiecke i und j in die gleiche Gruppe eingeordnet wurden	140
Abb. 5.10: Übereinstimmende Gruppierungen von mehreren Versuchspersonen	141
Abb. 5.11: Klassifizierungen derselben Person mit 3 und 8 Klassen	142
Abb. 5.12: Versuchsobjekt im Versuch 5. Die Farben sind durch die Muster  (schwarz), ::::: (rot) und  (blau) ersetzt.	143
Abb. 5.13: Versuchsobjekte im Versuch 5;  ^schwarz, ::::: ^rot,  ^blau	144
Abb. 5.14: Häufigkeitsverteilung der Gruppenzahlen in Versuch 5	145
Abb. 5.15: Modell des Merkmalsraumes von der Seite (a) und von oben (b) mit markierter Klasseneinteilung in (b)	146
Abb. 5.16: Modell des Merkmalsraumes aus der Sicht des Ursprungs des Koordinatensystems mit markierter Klasseneinteilung	147
Abb. 5.17: Klasseneinteilung aus Abb. 5.16 im transformierten Merkmalsraum	149
Abb. 5.18: Modell des Merkmalsraumes im Versuch 6 aus zwei verschiedenen Perspektiven	150

VERZEICHNIS DER WICHTIGSTEN SYMBOLE

$\mathbb{N}$	Menge der natürlichen Zahlen
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{R}^+$	Menge der positiven reellen Zahlen
$\mathbb{R}^m$	Menge der Zeilenvektoren mit m reellen Komponenten
$\mathbb{R}_m$	Menge der Spaltenvektoren mit m reellen Komponenten
$\mathbb{R}^{(n,m)}$	Menge der Matrizen mit n Zeilen und m Spalten und reellen Komponenten
$ x $	mit $x \in \mathbb{R}$, Absolutbetrag auf $\mathbb{R}$
$\mathbf{x}, \mathbf{y}$	$\in \mathbb{R}^m$, Zeilenvektoren
$\ \mathbf{x}\ $	mit $\mathbf{x} \in \mathbb{R}^m$, euklidische Norm des $\mathbb{R}^m$
$\mathbf{A}$	$\in \mathbb{R}^{(n,m)}$, $n \times m$ Matrix
$\mathbf{A}^t$	mit $\mathbf{A} \in \mathbb{R}^{(n,m)}$, Transponierte von $\mathbf{A}$
$\mathbf{A}^{-1}$	mit $\mathbf{A} \in \mathbb{R}^{(n,m)}$, Inverse von $\mathbf{A}$
$ \mathbf{A} $	mit $\mathbf{A} \in \mathbb{R}^{(n,m)}$, Determinante von $\mathbf{A}$
$\mathbf{I}_m$	$\in \mathbb{R}^{(m,m)}$, m-reihige Einheitsmatrix
$d(\mathbf{x}, \mathbf{y})$	mit $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, Distanzmaß auf dem $\mathbb{R}^m$
O_i	Objekt i, $1 \leq i \leq n$
O	$= \{O_1, \dots, O_n\}$, Objektmenge
I	$= \{1, \dots, n\}$, zu O gehörige Indexmenge
A_j	$\subseteq I$, Cluster j, $1 \leq j \leq k$
$\mathcal{P}(I)$	Potenzmenge von I
$\mathcal{G}$	$\subseteq \mathcal{P}(I)$, Gruppierung von I
$\mathcal{A}$	$= \{A_1, \dots, A_k\} \subset \mathcal{P}(I)$, Partition von I
$\mathcal{Z}$	$= (\mathcal{A}^1, \dots, \mathcal{A}^l)$, Hierarchie von Partitionen von I
n	Anzahl der Objekte
n_j	$= A_j $, Anzahl der Objekte in Cluster j, $1 \leq j \leq k$
m	Anzahl der Merkmale bzw. Variablen
k	Anzahl der Cluster

x_j	Merkmal j , $1 \leq j \leq m$, eindimensionale Zufallsvariable
$\mathbf{x}$	$= (x_1, \dots, x_m)$, m -dimensionale Zufallsvariable
$\mathbf{x}_i$	zu Objekt i , $1 \leq i \leq n$, gehörige m -dimensionale Zufallsvariable mit der Realisation $\mathbf{x}_i$
x_{ij}	$\in \mathbb{R}$, der an Objekt i beobachtete Wert des Merkmals j
$\mathbf{x}_i$	$= (x_{i1}, \dots, x_{im}) \in \mathbb{R}^m$, Vektor der an Objekt i beobachteten Merkmalswerte, Realisation von $\mathbf{x}_i$
$\mathbf{X}$	$= \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix} \in \mathbb{R}^{(n,m)}$, Datenmatrix
$\bar{x}_j$	$\in \mathbb{R}$, arithmetisches Mittel der Werte x_{ij} des Merkmals j , $1 \leq j \leq m$
$\bar{\mathbf{x}}$	$= (\bar{x}_1, \dots, \bar{x}_m) \in \mathbb{R}^m$, arithmetisches Mittel der Vektoren $\mathbf{x}_i$
$\bar{x}_{jl}$	$\in \mathbb{R}$, arithmetisches Mittel der Werte x_{ij} des Merkmals j bei Objekten aus Cluster l , $1 \leq j \leq m$, $1 \leq l \leq k$
$\mathbf{x}_l$	$= (\bar{x}_{l1}, \dots, \bar{x}_{lm}) \in \mathbb{R}^m$, arithmetisches Mittel der Vektoren $\mathbf{x}_i$ von Objekten aus Cluster l , $1 \leq l \leq k$
s_j^2	$\in \mathbb{R}^+$, Varianz der Werte des Merkmals j , $1 \leq j \leq m$
d_{ik}	Distanz zwischen Objekt i und k , $1 \leq i, k \leq n$
$\mathbf{D}$	$= (d_{ik})_{1 \leq i, k \leq n} \in \mathbb{R}^{(n,n)}$, Distanzmatrix
$\tilde{d}_{jl}$	Distanz zwischen Cluster j und l , $1 \leq j, l \leq k$
s_{ik}	Ähnlichkeit zwischen Objekt i und k , $1 \leq i, k \leq n$
$\mathbf{S}$	$= (s_{ik})_{1 \leq i, k \leq n} \in \mathbb{R}^{(n,n)}$, Ähnlichkeitsmatrix
$\mathbf{T}$	$\in \mathbb{R}^{(m,m)}$, Totale Streuungsmatrix
$\mathbf{B}$	$\in \mathbb{R}^{(m,m)}$, Zwischengruppenstreuungsmatrix
$\mathbf{W}$	$\in \mathbb{R}^{(m,m)}$, Innergruppenstreuungsmatrix
$\mathbf{W}_j$	$\in \mathbb{R}^{(m,m)}$, Totale Streuungsmatrix des Clusters j , $1 \leq j \leq k$
λ_i	$\in \mathbb{R}$, $1 \leq i \leq m$, Eigenwert von $\mathbf{W}^{-1}\mathbf{B}$

μ_j	$\in \mathbb{R}$, Erwartungswert der Variablen j , $1 \leq j \leq m$
$\boldsymbol{\mu}$	$= (\mu_1, \dots, \mu_m) \in \mathbb{R}^m$, Erwartungswert-(Mittelwert-)vektor der m -dimensionalen Zufallsvariablen $\mathfrak{X}$
$\boldsymbol{\mu}_j$	$\in \mathbb{R}^m$, Erwartungswertvektor der zu Cluster j , $1 \leq j \leq k$, gehörigen Zufallsvariablen $\mathfrak{X}_i$
σ_j^2	$\in \mathbb{R}^+$, Varianz der Variablen X_j , $1 \leq j \leq m$
Σ	$\in \mathbb{R}^{(m,m)}$, Kovarianzmatrix der m -dimensionalen Zufallsvariablen $\mathfrak{X}$
Σ_j	$\in \mathbb{R}^{(m,m)}$, Kovarianzmatrix der zu Cluster j , $1 \leq j \leq k$, gehörigen Zufallsvariablen $\mathfrak{X}_i$
$N_m(\boldsymbol{\mu}, \Sigma)$	m -dimensionale Normalverteilung mit Erwartungswertvektor $\boldsymbol{\mu}$ und Kovarianzmatrix Σ
$\mathfrak{X} \sim N_m(\boldsymbol{\mu}, \Sigma)$	Die Zufallsvariable $\mathfrak{X}$ unterliegt einer $N_m(\boldsymbol{\mu}, \Sigma)$ -Verteilung
$\hat{\cdot}$	Schätzwert

1. EINFÜHRUNG

1.1 CLUSTERANALYSE ALS MULTIVARIATES ANALYSEVERFAHREN

In vielen Situationen ist es sinnvoll, eine große Menge verschiedenartiger Objekte in kleinere, homogenere Teilmengen zu zerlegen. In den beiden folgenden Beispielen wird dies deutlich.

Beispiel 1:

Die Herstellung von Bekleidung¹ konnte durch die Idee, die vielen verschiedenen Kombinationen von Körpermaßen zu relativ wenigen Typen zusammenzufassen, wesentlich vereinfacht und dadurch kostengünstiger gestaltet werden. Maßkleidung, die genau für die speziellen Maße jeder einzelnen Person angefertigt wird, ist heute so teuer geworden, daß sich nur noch wenige Konsumenten diesen Luxus leisten können. Die Tatsache jedoch, daß sich viele Personen in ihrem Körperbau nur so wenig unterscheiden, daß sich die Unterschiede im Sitz der Kleidung kaum bemerkbar machen, ermöglicht es, für viele Personen dasselbe Schnittmuster zu verwenden. Dadurch brauchen nur noch wenige Typen - die Konfektionsgrößen - hergestellt werden, die in großen Serien und damit erheblich billiger gefertigt werden können als Maßkleidung. Dabei muß jedoch in Kauf genommen werden, daß die Konfektionskleidung den meisten Leuten nicht ganz genau paßt und manchen überhaupt nicht, weil sie einen untypischen Körperbau aufweisen.

Beispiel 2:

Betritt man auf der Suche nach einer speziellen Information z.B. eine große Universitätsbibliothek, so steht man mögli-

1 vgl. das Beispiel von STANGE, K.: Angewandte Statistik, 2. Teil, Mehrdimensionale Probleme. Berlin, Heidelberg, New York 1971, S. 85 ff. und BOCK, H.H.: Automatische Klassifikation. Göttingen 1974, S. 14 f.

cherweise einer Menge von weit über einer Million Bänden gegenüber. Obwohl die Wahrscheinlichkeit dafür, daß die gerade gesuchte Information darin enthalten ist, mit der Grösse des Buchbestandes zunimmt, kann ein so umfassendes Angebot gerade wegen seiner Grösse kaum noch ausgenutzt werden, wenn der Gesamtbestand nicht mittels eines ausgefeilten Systems so gegliedert wird, daß der Benutzer anhand der ihm bekannten Merkmale der gesuchten Veröffentlichung - das mögen der Autor, das Erscheinungsjahr, ein Stichwort oder ähnliches sein - bis in die unmittelbare Nähe des gewünschten Bandes geführt wird. Da man sich gerade bei wissenschaftlicher Arbeit häufig an inhaltlichen Merkmalen orientiert, hat es sich als zweckmäßig erwiesen, den Buchbestand außer nach äußerlichen Merkmalen wie Autorennamen, Titel u.ä. auch anhand der Ähnlichkeit der Inhalte zu strukturieren und in Sachgebiete zu zerlegen. Unterteilt man jedoch einen sehr großen Gesamtbestand in eine überschaubare Anzahl von Gebieten, so sind diese zwangsläufig meist selbst noch sehr umfangreich und müssen daher u.U. in mehreren Stufen in immer kleinere und dadurch zunehmend homogenere Teilgebiete zerlegt werden, bis diese so klein geworden sind, daß sie als Ganzes überschaubar sind.

Situationen wie in den vorangegangenen Beispielen kommen sowohl im Alltagsleben als auch im Bereich vieler - besonders empirischer - Wissenschaften häufig vor; deshalb wird der Technik, Objektmengen systematisch zu unterteilen (zu klassifizieren; Klassifikation) zunehmend mehr Aufmerksamkeit gewidmet.

Eine Objektmenge kann verhältnismäßig leicht in homogene Teilmengen zerlegt werden, solange die Menge nicht allzu viele Objekte enthält oder nur wenige Merkmale für die Unterscheidung der Objekte zugrundegelegt werden. Bei größeren Objekt- und Variablenzahlen wird es jedoch immer schwieriger, alle Objekte und Variablen *gleichzeitig* zu berücksich-

tigen. Zwar ist es möglich, bei einer großen Anzahl von Objekten zunächst immer nur einen Teil der gesamten Objektmenge zu klassifizieren und anschließend die gebildeten Klassen zu einer Klassifizierung der gesamten Objektmenge zusammenzufügen. Die Kombination der Klassen, die bei Betrachtung von Teilmengen als gut erscheinen mögen, führt aber nicht notwendigerweise auch zu einer angemessenen Klassifizierung der Gesamtmenge. Auch bei einer großen Zahl von Variablen ist es z.B. möglich, diese hintereinander zu verarbeiten, indem die eine zur Grobeinteilung und eine andere zur Feinteilung usw. benutzt werden; dabei wird aber auf jeder Stufe immer nur ein Teil der zur Verfügung stehenden Information verwendet, was in der Regel nicht zu einer bestmöglichen Klassifizierung führen kann.

Aufgrund dieser Schwierigkeiten wurden in den letzten Jahren¹ zahllose Vorschläge veröffentlicht, Klassifikationsprozesse so weit zu formalisieren, daß sie programmiert und folglich mit Hilfe von Computern durchgeführt werden können. Derartige Techniken zur Klassifikation wurden jedoch in mehreren Wissenschaftszweigen parallel entwickelt, sodaß es zu einem Wirrwarr von Bezeichnungen wie Clusteranalyse, Numerische oder Automatische Klassifikation, Taxonometrie, Numerische Taxonomie, Q-Analysis, Typologie, Clumping, Grouping, Unsupervised Learning usw. kam. Im Rahmen dieser Arbeit soll jedoch für alle Techniken zur Zerlegung einer Objektmenge in homogene Teilmengen einheitlich die weit verbreitete Bezeichnung "Clusteranalyse" (engl. "Cluster Analysis") verwendet werden.

Clusteranalysemethoden sind in den verschiedensten Situationen mit sehr unterschiedlichen konkreten Zielen konstruiert worden. Dementsprechend wurden auch von den Autoren die unterschiedlichsten Forderungen an den Ablauf und die Eigenschaften einer sinnvollen Clusteranalyse gestellt, sodaß heu-

1 Obwohl erste Ansätze schon 1939 bei TRYON, R.C.: Cluster Analysis. Ann Arbor 1939, zu finden sind, setzte die Entwicklung verstärkt erst nach 1970 ein. Vgl. hierzu STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. Berlin, New York 1977, S. 13.

te eine kaum übersehbare Vielfalt von Algorithmen und Programmen zur Verfügung steht¹. Diese lassen sich zwar auf vergleichsweise wenige grundsätzlich verschieden Vorgehensweisen zurückführen, durch mehr oder weniger große Abwandlung einzelner Schritte sind jedoch zahlreiche Varianten der dahinter oft nur noch schwer erkennbaren Prinzipien möglich. "Clusteranalyse" ist also im Gegensatz zu einigen anderen Analysemethoden wie z.B. der Regressionsanalyse nicht eine Bezeichnung für ein mathematisch-statistisches Verfahren, das durch die Angabe eines Modells, einer operationalen Fragestellung und eines eindeutigen Lösungsalgorithmus mit einer daraus resultierenden, scharf umrissenen Aussagekraft der Ergebnisse präzise definiert wird; vielmehr ist "Clusteranalyse" ein Sammelbegriff für größtenteils heuristische, teilweise auch von statistischen und entscheidungstheoretischen Überlegungen geprägte Verfahren, die - unabhängig vom konkreten Anwendungsbezug, eventuellen Modellvorstellungen und der algorithmischen Durchführung - die folgende Zielvorstellung gemeinsam haben:

Eine Menge von Objekten, die durch die Ausprägungen einer Anzahl von Merkmalen charakterisiert werden, soll so in Klassen zerlegt werden, daß die zu einer Klasse gehörigen Objekte möglichst ähnlich und die Klassen untereinander möglichst unähnlich sind².

1 Selbst wenn man nur die kommerziellen Marktforschungsinstitute in Deutschland in Betracht zieht, werden von diesen allein schon mehr als ein Dutzend verschiedener Systeme verwendet, die aufgrund unterschiedlicher Ergebnisse auch nur schwer miteinander vergleichbar sind. Eine ausführliche Gegenüberstellung findet sich bei WENDT, F.: Typologische Analyse Systeme in der deutschen Marktforschung. Hamburg, o.J.

2 vgl. z.B. BOCK, H.H.: Automatische Klassifikation. a.a.O., S. 13 f., VOGEL, F.: Probleme und Verfahren der numerischen Klassifikation. Göttingen 1975, S. 1. oder DURAN, B.S., ODELL, P.L.: Cluster Analysis, A Survey. Lecture Notes in Economics and Mathematical Systems, Vol. 100, Berlin, Heidelberg, New York 1974, S. 1.

Obwohl diese Zielsetzung als definierende Eigenschaft von Clusteranalyseverfahren einigermaßen unscharf formuliert ist und verschiedene Interpretationsmöglichkeiten zuläßt, kann man daran klar erkennen, daß das Clusteranalyseproblem ein grundsätzlich anderes ist als das Diskriminationsproblem oder das Identifikationsproblem.

Bei der Clusteranalyse (s. den oberen Teil von Abb. 1.1) wird eine Objektmenge, von der zunächst keine Gruppenstruktur bekannt ist (linkes Diagramm), erst in Klassen (Gruppen, Cluster o.ä.) unterteilt (Cluster 1 (▼) und Cluster 2 (■) im rechten Diagramm).

Dagegen ist die Aufgabe der Diskrimination (mittlerer Teil von Abb. 1.1) die Trennung bereits vorgegebener Gruppen (Gruppe 1 (▼) und Gruppe 2 (□) im linken Diagramm). Dazu wird ein Merkmal, eine Merkmalskombination oder eine aus den Merkmalsausprägungen eines jeden Objektes zu berechnende Funktion, die sog. Diskriminanzfunktion (s. rechtes Diagramm) gesucht, die es gestattet, anhand dieser Kenngröße möglichst viele Objekte den richtigen Gruppen zuzuordnen.

Bei der Identifikation wiederum (unterer Teil von Abb. 1.1) existieren zwar bereits bekannte Gruppen (Gruppe 1 (▼) und Gruppe 2 (■) im linken Diagramm), von einigen weiteren Objekten ist jedoch nicht bekannt, zu welcher Gruppe sie gehören (im linken Diagramm mit O gekennzeichnet). Die Aufgabe besteht nun darin, diese Objekte anhand ihrer Merkmalsausprägungen einer der beiden Gruppen zuzuordnen (s. rechtes Diagramm).

Die grundsätzliche Unterscheidung von Clusteranalyse, Diskrimination und Identifikation¹ schließt aber nicht aus,

1 Leider wurden die drei zu unterscheidenden Vorgänge und Verfahren nicht einheitlich mit denselben Bezeichnungen versehen. Besonders häufig wird "Klassifikation" auch oder sogar ausschließlich für die Identifikation im o.a. Sinne gebraucht. Eine ausführlichere Diskussion der Begriffe, allerdings mit starkem Bezug zur Biologie als Anwendungsgebiet, findet sich bei SNEATH, P.H.A., SOKAL, R.R.: Numerical Taxonomy. San Francisco 1973, S. 2 ff.

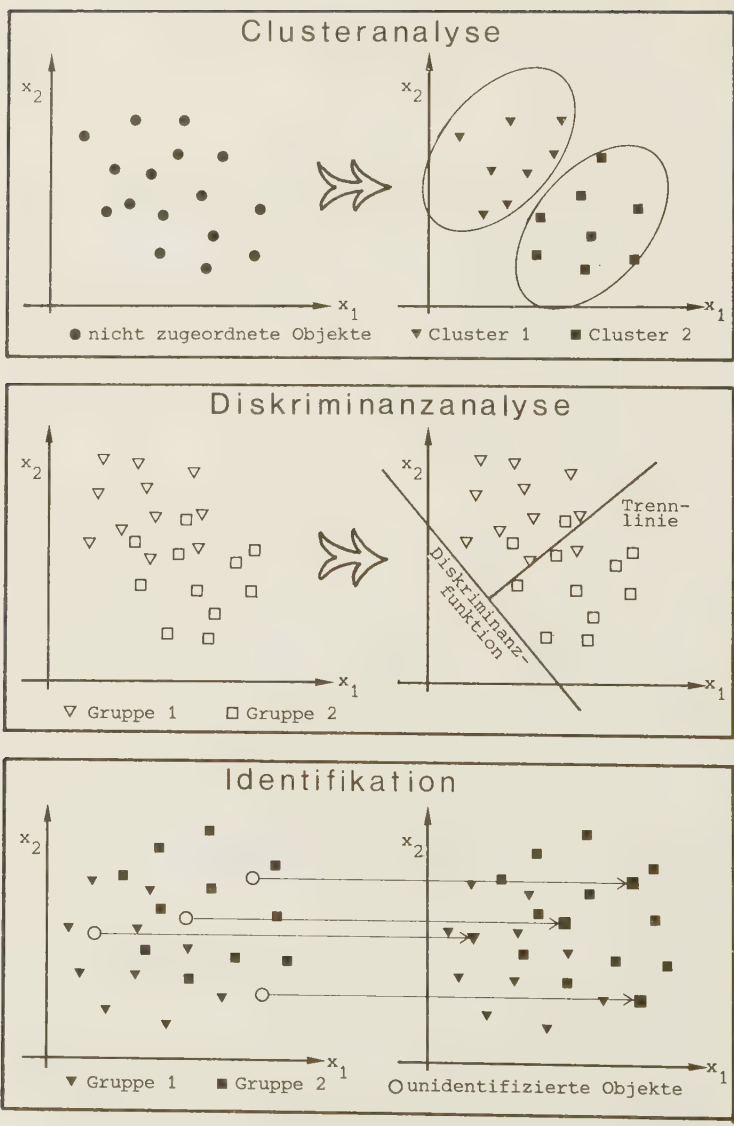


Abb. 1.1: Beispiele zur Wirkung von Clusteranalyse, Diskriminanzanalyse und Identifikation

daß diese drei Probleme, die ja oft gleichzeitig zu bewältigen sind, zueinander in Beziehung gesetzt werden können. Einerseits wird nämlich eine besonders deutliche Gruppenstruktur auch eine einfache Diskrimination ermöglichen, von daher wäre es durchaus plausibel, die Güte einer Clusteranalyse daran zu messen, wie gut die gebildeten Gruppen in einer Diskriminanzanalyse wieder voneinander getrennt werden können. Andererseits wird in vielen Situationen eine Diskriminanzanalyse, die ja eine vorgegebene Gruppenstruktur benötigt, auf den Ergebnissen einer Clusteranalyse aufbauen. Schließlich kann ein noch nicht einer bestimmten Gruppe zugeordnetes Objekt oft am besten mit Hilfe von Diskriminanzfunktionen identifiziert werden.

Neben der Reduktion der Daten durch Zusammenfassung ähnlicher Objekte zu Clustern ist die Verminderung der Variablenzahl ein anderer Weg, die in umfangreichen Datenmengen enthaltene Information zu komprimieren und transparenter zu machen. Die sogenannte Faktorenanalyse¹ geht davon aus, daß hinter den vielen beobachtbaren Variablen häufig nur relativ wenige stochastisch unabhängige Einflußgrößen stecken, die aber zunächst nicht erkennbar und u.U. auch selbst nicht meßbar sind. Durch algebraische Methoden kann aber immerhin die Anzahl dieser als Faktoren oder auch Supervariablen bezeichneten Größen und ihr Einfluß auf die einzelnen beobachteten Variablen festgestellt werden. Anhand der Korrelation zwischen den Faktoren und den Variablen können die Faktoren dann möglicherweise noch mit bestimmten Variablen identifiziert und inhaltlich beschrieben werden, obwohl sie von der Entstehung her zunächst reine Artefakte sind. Faßt man die stochastische Abhängigkeit bzw. hohe Korrelation zwischen Variablen als Ähnlichkeit auf, so ist die Faktorextraktion nichts anderes als das Ersetzen von

1 vgl. die Einführung bei ÜBERLA, K.: Faktorenanalyse. Berlin, Heidelberg, New York, 2. Aufl. 1971, S. 3.

vielen, teilweise ähnlichen Variablen durch wenige, möglichst unähnliche, d.h. stochastisch unabhängige Faktoren, die jeweils mehrere Variable repräsentieren. Insofern ist die Faktorenanalyse ein Analogon zu Clusteranalyse¹.

Bei der Verarbeitung großer Datenmengen, für die auch moderne Rechenanlagen nicht immer genug Speicherplatz bieten, benutzt man gelegentlich gleich beide Verfahren zur Datenreduktion. Beispielsweise kann zuerst mit Hilfe einer Faktorenanalyse die große Zahl der Variablen durch eine geringere Zahl von Faktoren ersetzt werden um anschließend durch Klassifikation die - nun bereits nur noch durch die Faktorenwerte beschriebenen - Objekte zu wenigen Typen zusammenzufassen. Allerdings ist die nachfolgende Charakterisierung der Klassen mit den in der Clusteranalyse verwendeten Faktoren nicht immer möglich bzw. sinnvoll, weil die Faktoren u.U. nicht inhaltlich interpretiert werden können. In diesem Fall kann man jedoch ohne weiteres zum Zwecke der Interpretation zu den ursprünglichen Variablen zurückkehren, obwohl die Clusteranalyse mit den Faktoren anstelle der Variablen durchgeführt wurde, denn bis auf einen geringen Rest enthalten ja Faktoren und Variablen die gleichen Informationen.

1 Die Analogie ist jedoch keinesfalls vollständig, d.h. Clusteranalyse und Faktorenanalyse lassen sich nicht durch Vertauschung von Objekten und Variablen ineinander überführen. Bei der Clusteranalyse ergibt sich z.B. die Ähnlichkeit der Cluster erst im Laufe des Verfahrens, während die Faktorenanalyse die stochastische Unabhängigkeit der Faktoren von vornherein erzwingt. Dadurch bedingt kann ein Objekt bei der Clusteranalyse genau einem Typ zugeordnet werden, wogegen die Variablen in der Regel mit mehreren, eventuell sogar allen Faktoren korreliert sind. Rein technisch gesehen, benötigt man für eine Faktorenanalyse auch mehr Objekte als Variable, was eine Analyse der Objektstruktur mit faktorenanalytischen Mitteln meist unmöglich macht.

1.2 ZIELE DER CLUSTERANALYSE

Die große Zahl der aus verschiedenen Anwendungsbereichen stammenden Verfahren zur Clusteranalyse stellt den Anwender vor das immer schwerer zu lösende Problem, ein für seine spezielle Aufgabenstellung geeignetes Verfahren auszuwählen. Zwar wird das Spektrum der Präzedenzfälle, in denen das eine oder andere Clusteranalyseverfahren zu sinnvollen Ergebnissen geführt hat, ständig breiter, aber die Verfahrensauswahl ist in den Beispielsituationen kaum jemals aufgrund zwingender Kriterien getroffen worden. Oft wird sie gar nicht und häufig nur mit Plausibilitäts-erklärungen begründet; in sehr vielen Untersuchungen sind in Ermangelung von Auswahlkriterien Clusteranalysen mit mehreren Verfahren durchgeführt worden, sodaß von den verschiedenen Ergebnissen das am leichtesten zu interpretierende verwendet werden konnte. Diese Vorgehensweise kann jedoch bestenfalls als Notlösung gelten, denn es ist ziemlich gleichgültig, welche Lösung man wählt, "Argumente für die ausgewählte Lösung lassen sich problemlos finden, vor allem bei entsprechend weicher, aber wohlklingender Formulierung"¹. Insbesondere begibt man sich auf diese Weise der wesentlichen Vorzüge der numerischen Klassifikation, größere Variablen- und Objektzahlen verarbeiten zu können, weil der menschliche Betrachter die unter Berücksichtigung aller Variablen und Objekte gewonnenen Ergebnisse nur in Bezug auf wenige Variable und Objekte beurteilen kann. Daher wäre es dringend notwendig, weniger subjektive Auswahlkriterien zu entwickeln, die vor allem vor der Durchführung der Verfahren herangezogen werden können. Leider ist die Zahl der Versuche, Clusteranalyse-

1 Aus der Stellungnahme von ISBA zum Vergleich der Typologiesysteme, die in der deutschen Marktforschung verwendet werden, in: WENDT, F.: Typologische Analyse Systeme in der deutschen Marktforschung. a.a.O., S. 117.

verfahren zu vergleichen und zu bewerten, noch sehr gering¹. Dies mag daran liegen, daß das Clusteranalyseproblem mit dem o.a. Ziel² nur unvollkommen definiert ist und ein und dasselbe Verfahren unter verschiedenen Aspekten als verschieden leistungsfähig eingestuft werden kann. Die meisten Vergleichsvorhaben sind deshalb auch mit empirischen Objektmengen unter bestimmten Gesichtspunkten vorgenommen worden und können folglich nur schwer verallgemeinert werden.

Auch das an sich sehr sinnvolle Vorhaben, Clusteranalyseverfahren in einer Methodenbank³ zu sammeln, die dem Benutzer gleichzeitig Entscheidungshilfen anbietet, krankt einstweilen einfach daran, daß als zwingendes Auswahlkriterium lediglich das Skalenniveau der Merkmale zur Verfügung steht, ansonsten aber lediglich Empfehlungen gegeben werden können, die zur eindeutigen Festlegung aber bei weitem nicht ausreichen. Man kennt zwar einige allgemeine Eigenschaften von bestimmten Methoden, wie z.B. die Tendenz, gleichgroße Gruppen von Kugelgestalt oder längere Ketten zu bilden. Eine Entscheidung aufgrund solcher Verfahrenscharakteristika bedeutet aber, sich gleichzeitig mit der Methode auf Ergebnisse mit bestimmten Eigenschaften festzulegen, was nur dann gerechtfertigt ist, wenn man eine diesbezügliche Vorinformation besitzt. In anderen Fällen sind solche Entscheidungshilfen nicht nur wenig nützlich, sondern sogar gefährlich, weil sie der Möglichkeit Raum geben, der zu klassifizierenden Objektmenge eine Struktur aufzuprägen, die

1 vgl. DURAN, B.S., ODELL, P.L.: Cluster Analysis, A Survey. a.a.O., S. 107

2 s.S. 5

3 An der Universität Augsburg wird gegenwärtig das Methodenbanksystem ANCA entwickelt. Vgl. SCHALLER, K.: Eine Methodenbank zur Clusteranalyse. In: BOCK, H.H. (Hrsg.): Klassifikation und Erkenntnis III. Frankfurt 1979, S. 79 - 94.

wesentlich von den subjektiven Vorstellungen des Anwenders und weniger von den Daten bestimmt wird.

Eine sinnvolle Bewertung von Verfahren wird solange nicht möglich sein, wie der mit einer Clusteranalyse verfolgte Zweck nicht präzisiert wird, weil erst dann, wenn das Ziel durch eine berechenbare Größe ausgedrückt werden kann, die Ergebnisse verschiedener Verfahren an diesem Ziel gemessen werden können. Wie verschieden aber die mit Clusteranalysen verfolgten Absichten sind, und wie schwer diese Absichten u.U. operationalisiert werden können, mögen die folgenden zwei Beispiele zeigen.

Beispiel 1: Das multiple Standortproblem¹

Eine Reihe von n Kunden in den Orten $x_1, \dots, x_n$ soll von m Auslieferungslagern (mit $m < n$) aus beliefert werden. Diese Auslieferungslager sollen in noch festzulegenden Orten $\tilde{x}_1, \dots, \tilde{x}_m$ errichtet werden, wobei die Orte so zu wählen sind, daß die Transportkosten insgesamt minimal werden. Die Kosten für den Transport von $\tilde{x}_j$ nach x_i werden als proportional zur geometrischen Entfernung $d(\tilde{x}_j, x_i)$ zwischen $\tilde{x}_j$ und x_i angesehen und werden wegen des unterschiedlichen Bedarfs der Kunden noch mit einem Faktor w_i gewichtet.

Die Menge $M = \{x_1, \dots, x_n\}$ der zu beliefernden Kundenstandorte ist also zunächst in Teilmengen $A_1, \dots, A_m$ zu zerlegen, die von jeweils einem Auslieferungslager aus beliefert werden. Die Teilmengen sollen die Bedingungen

$$A_1 \cup A_2 \cup \dots \cup A_m = M$$

1 In Anlehnung an SPÄTH, H.: Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion. München, Wien 1975, S. 131 ff. Dort werden auch Algorithmen zur Lösung dieses Problems angegeben.

und

$$A_i \cap A_j = \emptyset \quad \text{für } i \neq j, \quad 1 \leq i, j \leq m,$$

erfüllen, d.h. jeder Kunde wird beliefert und kein Kunde soll von zwei Lägern aus beliefert werden. Für jede Teilmenge A_i kann dann der optimale Lagerstandort $\tilde{x}_i$ als Funktion von A_i , also $\tilde{x}_i = \tilde{x}_i(A_i)$, $1 \leq i \leq m$, bestimmt werden. Die Zerlegung ist also so vorzunehmen, daß die Kostenfunktion

$$Z(A_1, \dots, A_m) = \sum_{j=1}^m \sum_{x_i \in A_j} w_i d(\tilde{x}_j(A_j), x_i)$$

ein Minimum annimmt.

Beispiel 2: Die Iris-Daten¹

Bei ähnlichen Pflanzenarten wie z.B. den verschiedenen Iris-Arten ist es u.U. nicht möglich, diese durch Betrachtung einzelner Merkmale zu unterscheiden. So weisen die Exemplare der Species *Iris setosa* beispielsweise in den meisten Fällen zwar längere Kelchblätter als die Exemplare der *Iris versicolor* auf, aber eben nicht immer, so daß eine sichere Unterscheidung dadurch nicht möglich ist. Hat man also eine Population von z.B. 150 Pflanzen vor sich, von denen nicht bekannt ist, welcher Art sie angehören, so stellt sich die Frage, ob es möglich ist, durch gleichzeitige Analyse mehrerer Merkmale zum einen herauszufinden, wieviele Arten darin vertreten sind, und zum anderen, welche Pflanze welcher Spezies angehört. Die Aufgabe, die gesamte Population in die einzelnen Arten zu

1 R.A. Fisher benutzte 1936 die an je 50 Exemplaren der drei Species *Iris setosa*, *Iris versicolor* und *Iris virginica* gemessenen Werte von vier Merkmalen als Basis für eine Diskriminanzanalyse, da er die Artzugehörigkeit der Pflanzen ja schon kannte. Seitdem werden sie aber immer wieder als Prüfstein für Clusteranalyseverfahren benutzt, indem untersucht wird, wie gut das Verfahren die Artzugehörigkeit aus den Merkmalswerten erkennen kann. S. FISHER, R.A.: The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7 (1936-37), S. 179 - 188

zu zerlegen, kann jedoch nicht direkt als Optimierungsaufgabe mit einer Zielfunktion ausgedrückt werden. Daher muß hier zunächst der Versuch gemacht werden, mit allgemeinen, vom konkreten Anwendungsfall unabhängigen Überlegungen die gesuchte Gruppenstruktur aus den Merkmalswerten herauszufiltern.

Ein wesentlicher Unterschied zwischen diesen Beispielen ist darin zu sehen, daß im zweiten Fall durch die Zugehörigkeit der einzelnen Pflanzen zu einer Art bereits von vornherein eine Gruppenstruktur vorhanden ist, die durch eine Clusteranalyse lediglich aufgedeckt werden soll. Dagegen wird im ersten Fall die gesuchte Clusterstruktur erst dadurch definiert, daß sie ein Minimum der Kostenfunktion darstellt. Während also im Falle der Iris-Daten eine *immanente* Gruppenstruktur möglichst unverändert wiedergegeben werden soll, wird im Problem der Transportkostenminimierung den Daten *von außen* eine Gruppenstruktur aufgeprägt.

Lange Zeit hindurch wurde unter "Clusteranalyse" fast ausschließlich die Aufdeckung zwar vorhandener, aber nicht ohne weiteres erkennbarer Gruppenstrukturen verstanden und die Existenz solcher Gruppen geradezu als Voraussetzung für die erfolgreiche Durchführung einer Clusteranalyse angesehen¹. Inzwischen setzt sich jedoch die Auffassung durch, daß in vielen Anwendungen die Strukturerkennung gar nicht beabsichtigt ist, sondern vielmehr eine - vom Vorhandensein von Gruppen unabhängige - in irgendeinem Sinne zweckmäßige Aufteilung der Objektmenge gewünscht wird².

1 Diese Ansicht wurde noch 1975 von Vogel vertreten. Vgl. VOGEL, F.: Probleme und Verfahren der numerischen Klassifikation. a.a.O., S. 1

2 Vgl. z.B. EVERITT, B.S.: Unresolved Problems in Cluster Analysis. Biometrics 35 (1979), S. 177. Everitt bezeichnet die beiden o.a. Fälle als "classification" und "dissection", führt jedoch an, daß sie in der Praxis häufig nicht klar voneinander getrennt werden.

In Anlehnung an SPÄTH¹ sollen daher im Rahmen dieser Arbeit

- (a) die Aufgabe, in der Objektmenge vorhandene Cluster zu erkennen, als Strukturanalyse und
- (b) die Aufgabe, eine Objektmenge möglichst zweckmäßig zu zerlegen, ohne daß dabei die Existenz mehrerer Cluster vorausgesetzt bzw. das Erkennen der Cluster angestrebt wird, als Klassifikation

bezeichnet werden. Unabhängig von der Aufgabe, zu der sie jeweils eingesetzt werden, sollen jedoch alle Methoden zur Zerlegung einer Objektmenge in möglichst homogene Teilmengen weiterhin einheitlich als Clusteranalyseverfahren bezeichnet werden.

Je nachdem für welche der beiden o.g. Aufgaben ein Clusteranalyseverfahren eingesetzt werden soll, müssen unterschiedliche Anforderungen an das Verfahren gestellt werden:

Bei der Strukturanalyse hängt die Eignung eines Clusteranalyseverfahrens davon ab, ob es in der Lage ist, die folgenden drei Teilaufgaben zu lösen bzw. eine Entscheidungshilfe zu geben:

- Zunächst sollte in jedem Fall festgestellt werden können, ob die untersuchte Objektmenge *überhaupt mehrere Cluster* enthält oder nicht. Dabei tritt wie beim "Clusteranalyseproblem" selbst² die Schwierigkeit zutage, den Begriff "Cluster" ausreichend genau zu definieren und einen Indikator zu finden, der gestattet zu entscheiden, ob die

1 S. SPÄTH, H.: Cluster-Analyse: Konzepte, Rezepte und Anwendungen. Manuskript zu einem Vortrag beim 11. Steiermärkischen Mathematischen Symposium in Graz, 1.-4.10.1979, S. 1. Späth schlägt vor, die Untersuchung unbekannter Strukturen mit "Cluster-Analyse" und eine möglichst gute Aufteilung mit "automatischer Klassifikation" zu bezeichnen, behält aber trotzdem "Cluster-Analyse" als Oberbegriff für beide Problembereiche bei.

2 Vgl. S. 5 unten.

von einem Algorithmus gefundene Zerlegung eine vorgefundene Struktur wiedergibt oder ein methodisches Artefakt darstellt.

- Mit Clusteranalyseverfahren kann jede Objektmenge in praktisch jede beliebige Anzahl von Clustern zerlegt werden. Daher ist zu entscheiden, welche *Clusterzahl* der Clusterstruktur der Objektmenge entspricht.
- Durch die Anwendung mehrerer Verfahren können zu derselben Clusterzahl mehrere verschiedene Lösungen erzeugt werden. U.U. kann auch ein und dasselbe Verfahren - etwa durch Abhängigkeit von der Eingabereihenfolge der Objekte - verschiedene Lösungen mit derselben Anzahl von Clustern liefern. In diesem Fall muß auch noch eine Entscheidung zwischen den *verschiedenen Lösungen* getroffen werden.

Die Aufgabe der Klassifikation im o.a. Sinne besteht im Wesentlichen darin, ein Maß für die Zweckmäßigkeit einer Zerlegung festzulegen und dieses Maß durch einen geeigneten Clusteranalysealgorithmus zu optimieren. Hinsichtlich der Festlegung eines Maßes lassen sich zwei Fälle unterscheiden:

- Wenn aus der Anwendungssituation heraus wie in dem Multiplen Standortproblem¹ eine *Zielfunktion* abgeleitet werden kann, ist damit eindeutig geklärt, daß diejenige Zerlegung, die den besten Zielfunktionswert aufweist, auch im Sinne des Anwendungsproblems die beste Lösung ist. Obwohl das noch verbleibende Problem, diese Zielfunktion zu optimieren, durchaus nicht trivial ist, soll dieser Fall hier nicht weiter diskutiert werden, weil die Optimalität einer Lösung in dieser Situation bereits eindeutig definiert ist und keine diesbezüg-

¹ Vgl S. 11

liche Entscheidungsfreiheit bzw. Unsicherheit mehr zurückbleibt.

- Die relativ häufige Situation, daß die untersuchte Objektmenge weder mehrere Cluster im Sinne der Strukturanalyse¹ enthält noch eine Zielfunktion aus dem Anwendungsbezug abgeleitet werden kann, ist die bei weitem schwierigste. Die fehlende Zielfunktion hat zur Folge, daß verschiedene Lösungen i.a. nicht mehr in eine eindeutige Rangordnung gebracht werden können bzw. - formalmathematisch ausgedrückt - keine hinreichende Bedingung für die Optimalität einer Lösung mehr angegeben werden kann. Es ist jedoch immer noch möglich, stattdessen wenigstens einige *notwendige Bedingungen* anzugeben, die ein gutes Klassifikationsergebnis auf jeden Fall erfüllen sollte²; mit Hilfe derartiger Bedingungen, die beispielsweise die Invarianz gegenüber Skalentransformationen oder Stetigkeit³ beinhalten können, wird die Zahl der brauchbaren Verfahren u. U. stark eingeschränkt.

Dieser Gliederung der Clusteranalyse in die beiden Teilbereiche "Strukturanalyse" und "Klassifikation" entsprechend wird in einem Teil dieser Arbeit (Kapitel 4) unter-

-
- 1 In der Stellungnahme zum Vergleich der Typologiesysteme schätzt die COMPAGNON Marktforschungs-Institut GmbH, daß in der Praxis mindestens in 80 bis 90 von 100 Fällen "Datenstrukturen, die letztlich nur ein Dichtezentrum aufzeigen und in Richtung der Randbezirke immer dünner besiedelt sind", vorliegen. S. WENDT, F.: Typologische Analyse Systeme in der deutschen Marktforschung. a.a.O., S. 109
 - 2 In dieser Richtung sind bisher nur relativ wenige Untersuchungen angestellt worden, so z.B. von FISHER, L., van NESS, J.W.: Admissible Clustering Procedures. Biometrika 58 (1971), S. 91 - 104 und DEGENS, O.P.: Konsistenzforderung für die Automatische Klassifikation. In: BOCK, H.H. (Hrsg.): Klassifikation und Erkenntnis III. Frankfurt 1979, S. 31 - 47.
 - 3 Mit Stetigkeit eines Clusteranalyseverfahrens ist gemeint, daß die geringfügige Änderung eines Meßwertes das Ergebnis nicht wesentlich verändert. Vgl. DEGENS, O.P.: Konsistenzforderung für die automatische Klassifikation. a.a.O., S. 37

sucht, mit welchen numerischen Verfahren vorgegebene Strukturen in synthetischen Datensätzen am besten erkannt werden können, wobei die zu erkennenden Strukturen im Rahmen eines geeigneten Modells definiert werden. Ein weiterer Teil der Arbeit (Kapitel 5) wird den Eigenschaften gewidmet, die eine gute Klassifikation bzw. ein gutes Klassifikationsverfahren aufweisen sollte; diese werden hier nicht aus theoretischen Untersuchungen, sondern aus experimentellen Beobachtungen abgeleitet.

Zuvor sollen jedoch die verwendeten numerischen Verfahren vorgestellt werden (Kapitel 2) und ein Überblick über die Vielzahl der mit der Durchführung einer Clusteranalyse verbundenen Detailprobleme gegeben werden (Kapitel 3).

2 METHODEN DER CLUSTERANALYSE

2.1 ALLGEMEINES

Sei $O = \{O_1, O_2, \dots, O_n\}$ mit $n \in \mathbb{N}$ eine Menge von n Objekten, die mittels eines Clusteranalyseverfahrens untersucht werden sollen.

Werden an den Objekten mehrere Merkmale $X_1, \dots, X_m$, $m \in \mathbb{N}$, gemessen, so lassen sich die Beobachtungswerte in Form einer Datenmatrix

$$X = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nm} \end{pmatrix} = (x_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} \in \mathbb{R}^{(n,m)} \quad (2.1)$$

zusammenfassen, wobei x_{ij} die Ausprägung des Merkmals j beim Objekt i sein soll. Das Objekt i wird also durch den Zeilenvektor

$$x_i = (x_{i1} \dots x_{im}) = (x_{ij})_{1 \leq j \leq m} \in \mathbb{R}^m \quad (2.2)$$

beschrieben, während die Information über das Merkmal j in der j -ten Spalte von X enthalten ist.

Für einige Verfahren wird als Basis auch statt der Datenmatrix X eine Distanzmatrix

$$D = \begin{pmatrix} d_{11} & \dots & d_{1n} \\ \vdots & & \vdots \\ d_{n1} & \dots & d_{nn} \end{pmatrix} = (d_{ij})_{1 \leq i, j \leq n} \in \mathbb{R}^{(n,n)} \quad (2.3)$$

oder eine Ähnlichkeitsmatrix

$$S = \begin{pmatrix} s_{11} & \dots & s_{1n} \\ \vdots & & \vdots \\ s_{n1} & \dots & s_{nn} \end{pmatrix} = (s_{ij})_{1 \leq i, j \leq n} \in \mathbb{R}^{(n,n)} \quad (2.4)$$

benötigt. Darin ist d_{ij} ein Maß für die Unähnlichkeit - bei geometrischer Interpretation die Distanz - der Objekte i und j , welches nicht unbedingt metrische Eigenschaften haben muß¹. Immerhin wird aber vorausgesetzt, daß

$$d_{ij} \geq 0 \quad 1 \leq i, j \leq n \quad , \quad (2.5)$$

$$d_{ii} = 0 \quad 1 \leq i \leq n \quad , \quad (2.6)$$

$$d_{ij} = d_{ji} \quad 1 \leq i, j \leq n \quad (2.7)$$

gilt². Eine Distanzmatrix ist daher symmetrisch und enthält auf der Hauptdiagonalen nur Nullen. Analog ist s_{ij} ein Maß für die Ähnlichkeit - Nähe - der Objekte i und j , das die Bedingungen

$$s_{\min} \leq s_{ij} \leq s_{\max} \quad 1 \leq i, j \leq n \quad , \quad (2.8)$$

$$s_{ii} = s_{\max} \quad 1 \leq i \leq n \quad , \quad (2.9)$$

$$s_{ij} = s_{ji} \quad 1 \leq i, j \leq n \quad (2.10)$$

$$\text{mit } s_{\min} \leq s_{\max} \text{ und } s_{\min}, s_{\max} \in \mathbb{R}$$

einhält³. Distanz- und Ähnlichkeitsmatrizen lassen sich durch eine Reihe von Transformationen ineinander überführen.

Die Aufgabe der Clusteranalyseverfahren ist es nun, nach bestimmten Regeln bzw. Kriterien Teilmengen der Objekt-

1 Eine Metrik d ist eine Abbildung, die je zwei Punkten x und y eines Vektorraumes eine nichtnegative reelle Zahl $d(x,y)$ zuordnet, wobei die folgenden drei Bedingungen eingehalten werden:

$$d(x,y) = 0 \leftrightarrow x = y$$

$$d(x,y) = d(y,x) \quad (\text{Symmetrie})$$

$$d(x,y) \leq d(x,z) + d(z,y) \quad (\text{Dreiecksungleichung})$$

2 s. BOCK, H.H.: Automatische Klassifikation. a.a.O., S. 25

3 ebenda, S. 25

menge O zu bilden, oder, was dasselbe ist, Teilmengen der zu O gehörigen Indexmenge $I = \{1, \dots, m\}$ zu bilden. Diese Teilmengen werden als Cluster, Gruppen, Klassen o.ä. bezeichnet¹.

Das Endergebnis eines Clusteranalyseverfahrens ist jeweils ein System von Clustern:

Eine Gruppierung $\mathcal{G}$ ist eine Menge von nicht leeren Teilmengen $A_1, \dots, A_k$, $k \in \mathbb{N}$, der Indexmenge I , d.h.

$$\mathcal{G} = \{A_1, \dots, A_k\} \quad \text{mit } A_i \subseteq I, 1 \leq i \leq k \quad . \quad (2.11)$$

Die Gruppierung heißt disjunkt, wenn

$$A_i \cap A_j = \emptyset \quad \text{für } i \neq j, 1 \leq i, j \leq k \quad (2.12)$$

gilt; sie heißt exhaustiv, wenn

$$\sum_{i=1}^k A_i = I \quad (2.13)$$

gilt.

In den meisten Untersuchungen, die sich der Clusteranalyse bedienen oder Clusteranalyse selbst zum Gegenstand haben, werden fast ausschließlich exhaustive, disjunkte Gruppierungen verwendet:

Eine Partition α der Länge k ist eine disjunkte, exhaustive Gruppierung mit k Clustern.

¹ Als Cluster können sowohl die Teilmengen von O als auch die Teilmengen von I definiert werden. Wegen der eindeutigen Zuordnung von Objekt und Index macht dies inhaltlich keinen Unterschied. Für formale Zwecke ist es jedoch günstiger, eine Indexmenge als Cluster zu bezeichnen. Vgl. auch SPÄTH, H.: Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion. a.a.O., S. 30.

Eine Partition $\alpha = \{A_1, \dots, A_k\}$, $k \in \mathbb{N}$ heißt feiner als eine Partition $\beta = \{B_1, \dots, B_l\}$, $l \in \mathbb{N}$, $l < k$, wenn jedes Cluster von α vollständig in einem Cluster von β enthalten ist, d.h. wenn $\forall A_i \in \alpha \exists B_j \in \beta : A_i \subseteq B_j$.

Eine Hierarchie γ von Partitionen ist eine Folge $\alpha^1, \dots, \alpha^l$, $1 \leq l$, von Partitionen mit

$$\alpha^1 = \{\{1\}, \dots, \{n\}\}, \quad (2.14)$$

$$\alpha^i \text{ ist feiner als } \alpha^{i+1} \text{ f\"ur } 1 \leq i < l, \quad (2.15)$$

$$\alpha^l = \{I\}. \quad (2.16)$$

Die Verfahren der Clusteranalyse sind außerordentlich zahlreich und können nach verschiedenen Kriterien systematisiert werden:

Nach der Zahl der bei jeder Zuordnung berücksichtigten Variablen lassen sich monothetische und polythetische Verfahren unterscheiden: Ein monothetisches Verfahren berücksichtigt bei der Trennung von je zwei Gruppen immer nur eine Variable, während ein polythetisches Verfahren alle Variablen simultan verwendet. Monothetische Verfahren sind relativ einfach zu benutzen und haben deshalb in der Praxis, besonders dort, wo die Klassifikation manuell betrieben wird, sicher Vorteile. Wie das Beispiel in Abb. 2.1 zeigt, sind aber in vielen Fällen Gruppen aufgrund eines einzigen Merkmals nicht sauber zu trennen, weshalb monothetische Verfahren den polythetischen für Zwecke der Clusteranalyse sicher unterlegen sind und in dieser Arbeit nicht weiter betrachtet werden¹.

¹ Zu den monothetischen Verfahren gehört z.B. die AID-Methode (Automatic Interaction Detection), bei der zu jeder weiteren Unterteilung das dazu am besten geeignete Merkmal herausgesucht wird. Dadurch können Einsichten in die unterschiedliche Wichtigkeit der Variablen in den verschiedenen Gruppen gewonnen werden. Beispielsweise ist für die Gliederung der weiblichen Bevölkerung der Familienstand wichtig, während bei den Männern die Berufsausbildung dominierend ist. S. ANDERBERG, M.R.: Cluster Analysis for Applications. New York, London 1973, S. 195.

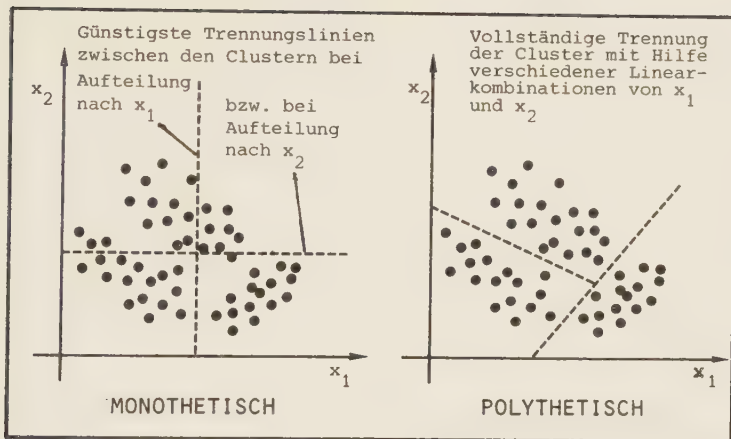


Abb. 2.1: Beispiel mit monothetischer und polythetischer Aufteilung

Ein anderes Unterscheidungsmerkmal für Clusteranalyseverfahren ist das Prinzip der Clusterbildung. Iterative Verfahren gehen von einer bereits gegebenen Startpartition aus, die z.B. durch Zufall erzeugt werden kann, und versuchen, durch Umordnung von Elementen eine Zielfunktion zu verbessern. Diese Versuche werden solange fortgesetzt, bis durch die Umordnung einzelner Elemente keine Verbesserung mehr erreicht werden kann. Dadurch ist aber nicht gesichert, daß die gefundene Partition bereits optimal ist, sondern nur, daß ein Suboptimum erreicht wurde. Das globale Optimum kann aber häufig dadurch gefunden werden, daß man ein Verfahren mehrfach mit unterschiedlichen Eingabereihenfolgen oder anderen Startpartitionen durchführt; dabei findet man i.d.R. verschiedene Suboptima, von denen das beste meist mehrfach vorkommt und vermutlich auch das globale Optimum ist¹.

1 Eigene Simulationsversuche haben ergeben, daß bei 2 bis 3 Clustern und 25 Objekten schon mit 3 verschiedenen Startpartitionen das globale Optimum des Varianzkriteriums meistens gefunden wird. Mit steigender Clusterzahl nimmt jedoch die Anzahl der notwendigen Versuche stark zu. Vgl. S. 118 ff.

Hierarchische Verfahren bilden jeweils Folgen von Partitionen, die dadurch entstehen, daß fortgesetzt bereits bestehende Cluster zusammengefaßt bzw. aufgespalten werden.

Agglomerative hierarchische Verfahren beginnen mit der Menge aller einelementigen Cluster und fassen nach bestimmten Kriterien Schritt für Schritt immer mehr Cluster zusammen, bis alle Elemente zu einem Cluster gehören. Divisive hierarchische Verfahren teilen schrittweise die bestehenden Cluster immer weiter auf, bis jedes Cluster nur noch ein Element enthält. Auf beiden Wegen entsteht eine Hierarchie, die am besten graphisch durch ein sog. Dendrogramm wie in Abb. 2.2 dargestellt werden kann.

Daraus kann man entnehmen, daß bei agglomerativer Vorgehensweise zunächst die Objekte O_2 und O_6 und dann die Objekte O_1 und O_5 zu je einem Cluster zusammengefügt werden. Danach werden das Objekt O_4 dem Cluster $\{O_2, O_6\}$ und später O_3 zum Cluster $\{O_2, O_4, O_6\}$ hinzugefügt. Zum Schluß werden die Cluster $\{O_1, O_5\}$ und $\{O_2, O_3, O_4, O_6\}$ zur Objektmenge O vereinigt. Bei divisiver Vorgehensweise wird umgekehrt zuerst O aufgeteilt in die Cluster $\{O_1, O_5\}$ und $\{O_2, O_3, O_4, O_6\}$ und anschlies-

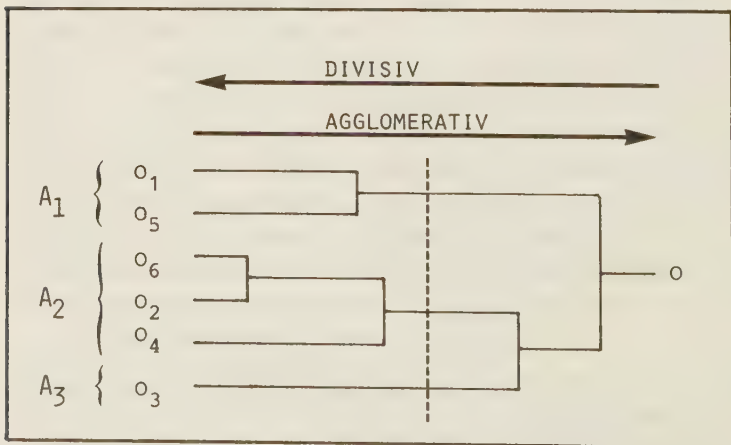


Abb. 2.2: Darstellung eines Dendrogramms bei agglomerativer und divisiver Zerlegung und Partition mit 3 Clustern

send werden nacheinander O_3 und O_4 aus dem Cluster $\{O_2, O_3, O_4, O_6\}$ herausgenommen. Zuletzt werden die Cluster $\{O_1, O_5\}$ und $\{O_2, O_6\}$ in die einzelnen Elemente zerlegt.

Eine Partition mit beispielsweise drei Clustern kann man aus dem Dendrogramm bestimmen, indem es an einer Stelle abgeschnitten wird, an der es gerade drei Äste aufweist (gestrichelte Linie in Abb. 2.2). Dadurch ergeben sich drei Teilbäume, an deren Endästen jeweils die zu einem Cluster (A_1, A_2 und A_3 in Abb. 2.2) gehörigen Objekte abgelesen werden können.

Da die Mehrzahl der divisiven Verfahren monothetisch arbeitet¹, und der Aufwand für polythetische divisive Verfahren erheblich größer ist als für agglomerative Verfahren², werden den divisive hierarchische Verfahren allgemein nur wenig benutzt und in diese Arbeit nicht mit einbezogen. Ferner werden hier - wie oben bereits angedeutet wurde - nur exhaustive disjunkte Verfahren untersucht, weil diese am weitesten verbreitet sind. Jedoch können auch einige gewichtige Argumente gegen die etwas einseitige Bevorzugung dieses Konzeptes vorgetragen werden:

- In der Marktforschung beispielsweise ist es vernünftig, nur Ballungen von Objekten als Typen zu betrachten und die dazwischen oder am Rande liegenden verstreuten Individuen in einen "Mülleimer" zu befördern³, da sie als eigenes Segment zahlenmäßig nicht stark genug bzw. zu heterogen sind.

1 s. VOGEL, F.: Probleme und Verfahren der numerischen Klassifikation. a.a.O., S. 244.

2 s. BOCK, H.H.: Automatische Klassifikation. a.a.O., S. 418.

3 Diese Ansicht wird z.B. von WENDT energisch vertreten. WENDT, F.: Typologische Analyse Systeme in der deutschen Marktforschung. a.a.O., S. 24.

- In einigen Anwendungen ist die Einordnung eines Objektes in mehrere Cluster, also eine nichtdisjunkte Gruppierung, vorteilhaft; z.B. sollte ein Kriminalroman, dessen Hauptperson ein Arzt ist, im Sachkatalog einer größeren Bibliothek sowohl unter "Kriminalromane" als auch unter "Arztromane" eingeordnet werden, weil das Wiederauffinden des Romans dadurch erleichtert wird.
- Neuere Entwicklungen deuten darauf hin, daß durch Clusteranalyse mit "unscharfen Mengen" (fuzzy sets) einige Probleme leichter als mit herkömmlichen Methoden gelöst werden können. Die Angabe eines Clusters in Form einer "unscharfen Menge" bedeutet, daß nicht angegeben wird, zu welchem Cluster ein Objekt gehört, sondern es wird für jedes Cluster und jedes Objekt angegeben, mit welcher Wahrscheinlichkeit bzw. mit welchem Anteil das Objekt zu dem betreffenden Cluster gehört. Anteile von 1 und 0 bedeuten wie bisher klare Zugehörigkeit bzw. Nichtzugehörigkeit. Anteile zwischen 0 und 1 deuten jedoch auf Unentscheidbarkeit, d.h. auf zwischen den "Kernen" der Cluster liegende Objekte oder auch Ausreißer hin¹.

Die Erfahrungen im Umgang mit solchen Verfahren sind aber offenbar noch vergleichsweise gering gegenüber den exhaustiven, disjunkten Verfahren, die sich in der Anwendung weitgehend durchgesetzt haben.

¹ s. BOCK, H.H.: Clusteranalyse mit unscharfen Partitionen. In: BOCK, H.H. (Hrsg.): Klassifikation und Erkenntnis III. Frankfurt 1979, S. 137 - 163.

2.2 HIERARCHISCHE VERFAHREN

Die in der vorliegenden Untersuchung verwendeten agglomerativen hierarchischen Clusteranalyseverfahren laufen alle nach dem folgenden Schema ab:

Ausgehend von der feinsten Partition, in der jedes Objekt ein eigenes Cluster bildet, werden in jedem Schritt die jeweils "ähnlichsten" Cluster zusammengefügt. Dabei nimmt die Länge der Partition ständig ab, bis im letzten Schritt nur noch ein einziges Cluster übrig bleibt. Die Unterschiede zwischen den Verfahren liegen nur in der verschiedenen Auffassung über die Ähnlichkeit bzw. Unterschiedlichkeit von Clustern¹.

- (1) Beim Single-Linkage-Verfahren wird die Distanz $\tilde{d}_{ij}$ zwischen zwei Clustern A_i und A_j lediglich durch die Distanz der beiden nächstgelegenen Elemente O_p aus A_i und O_q aus A_j definiert, d.h.

$$\tilde{d}_{ij} = \min \{d_{pq} \mid p \in A_i, q \in A_j\} \quad . \quad (2.17)$$

- (2) Das Complete-Linkage-Verfahren mißt die Unterschiedlichkeit zweier Cluster am Abstand der beiden am weitesten voneinander entfernten Elemente, also durch

$$\tilde{d}_{ij} = \max \{d_{pq} \mid p \in A_i, q \in A_j\} \quad . \quad (2.18)$$

¹ Die Namen der verschiedenen Verfahren werden leider in der Literatur nicht einheitlich verwendet. Es werden nicht nur einige weitere Bezeichnungen benutzt, sondern auch Namen vertauscht. Bei SNEATH, P.H.A., SOKAL, R.R.: Numerical Taxonomy. a.a.O., werden z.B. Average Linkage und Weighted Average Linkage genau Anders herum bezeichnet. Die hier gewählten Benennungen stammen aus STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 77.

- (3) Average-Linkage benutzt, quasi als Kompromiß zwischen Single-Linkage und Complete-Linkage, alle Distanzen zwischen Objekten aus A_i und A_j zur Berechnung des Abstandes von A_i und A_j . Dabei werden diejenigen Objekte, die zuletzt zu A_i bzw. A_j hinzugekommen sind, stärker gewichtet durch Berücksichtigung der Anzahl der Fusionschritte, die jedes Objekt bereits hinter sich hat. Bezeichnet man die Anzahl der Fusionsschritte, die die Objekte $O_p \in A_i$ bzw. $O_q \in A_j$ hinter sich haben, mit c_p bzw. c_q , so wird der Clusterabstand als

$$\tilde{d}_{ij} = \frac{1}{n_i n_j} \sum_{p \in A_i} \sum_{q \in A_j} \frac{1}{c_p + c_q} d_{pq} \quad , \quad (2.19)$$

berechnet, wobei

$$n_i = |A_i| \quad , \quad 1 \leq i \leq k \quad (2.20)$$

die Anzahl der Elemente von Cluster i ist.

- (4) Weighted Average-Linkage verwendet stattdessen die Zwischengruppendistanzen ohne Gewichtungsfaktoren, also

$$\tilde{d}_{ij} = \frac{1}{n_i n_j} \sum_{p \in A_i} \sum_{q \in A_j} d_{pq} \quad . \quad (2.21)$$

Die Verfahren (1) bis (4) benötigen als Informationsbasis lediglich eine Distanzmatrix $D = (d_{ij})$, die beliebig definierte Distanzen enthalten kann. Die nun folgenden Verfahren können zwar ebenfalls mit allen möglichen Distanzmaßen durchgeführt werden, jedoch ist bei den Verfahren (5), (6) und (7) eine für Interpretierbarkeit und bei (6) und (7) sogar für die Implementierung der Algorithmen wesentliche Voraussetzung, daß die vorgegebene Matrix D durch

$$d_{ij} = \sum_{l=1}^m (x_{il} - x_{jl})^2 \quad (2.22)$$

definierte sog. quadrierte euklidische Distanzen enthält. Durch die Anwendung der entsprechenden Algorithmen auf

andere Distanzen werden die Ergebnisse u.U. unbrauchbar und die einzelnen Fusionsschritte unplausibel.

- (5) Beim Median-Verfahren wird die Unterschiedlichkeit zweier Cluster A_i und A_j an der Distanz zwischen zwei Repräsentanten $\mathbf{x}_i^*$ und $\mathbf{x}_j^*$ mit $\mathbf{x}_i^*, \mathbf{x}_j^* \in \mathbb{R}^m$, gemessen:

$$\tilde{d}_{ij} = d(\mathbf{x}_i^*, \mathbf{x}_j^*) \quad . \quad (2.23)$$

Dabei wird ein Cluster mit nur einem Element von diesem Element selbst repräsentiert und bei der Vereinigung zweier Cluster i und j wird der Repräsentant $\mathbf{x}^*$ des neuen Clusters aus $\mathbf{x}_i^*$ und $\mathbf{x}_j^*$ als

$$\mathbf{x}^* = \frac{1}{2}(\mathbf{x}_i^* + \mathbf{x}_j^*) \quad (2.24)$$

berechnet.

- (6) Beim Centroid-Verfahren wird der Gruppenmittelwert (Centroid) $\bar{\mathbf{x}}_j$, also

$$\bar{\mathbf{x}}_j = \frac{1}{n_j} \sum_{i \in A_j} \mathbf{x}_i \quad (2.25)$$

als Vertreter des Clusters A_j zur Berechnung der Abstände

$$\tilde{d}_{ij} = d(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) \quad (2.26)$$

herangezogen, dadurch wird - anders als beim Median-Verfahren - implizit die unterschiedliche Mächtigkeit der Cluster berücksichtigt.

- (7) WARDS Verfahren liegt die Absicht zugrunde, das sog. Varianzkriterium

$$K_1(\mathbf{A}, \mathbf{X}) = \sum_{j=1}^k \sum_{i \in A_j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2 \quad (2.27)$$

als Maß für die Heterogenität der Cluster zu minimieren. Dabei ist

$$\|\mathbf{x}_i\| = \sqrt{\sum_{j=1}^m x_{ij}^2} \quad (2.28)$$

die euklidische Norm des $\mathbb{R}^m$ und daher

$$\|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{\sum_{p=1}^m (x_{ip} - x_{jp})^2} \quad (2.29)$$

gerade die euklidische Distanz von $\mathbf{x}_i$ und $\mathbf{x}_j$. Der durch die Fusion von A_i und A_j bewirkte Zuwachs des Varianzkriteriums, der nun wieder als Distanz interpretiert werden kann - es werden ja die Cluster mit dem kleinsten Zuwachs vereinigt - ergibt sich dann als

$$\tilde{d}_{ij} = \frac{n_i n_j}{n_i + n_j} \|\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j\|^2 \quad (2.30)$$

Mit Hilfe einer bekannten Formel von LANCE & WILLIAMS¹ lassen sich die Verfahren (1) bis (7) rekursiv definieren, sodaß die sich ständig verändernden Gruppenabstände $\tilde{d}_{ij}$ nicht immer von neuem aus der ursprünglichen, meist umfangreichen Distanzmatrix $\mathbf{D}$ oder sogar aus der Datenmatrix $\mathbf{X}$ berechnet werden müssen. Stattdessen wird bei jeder Fusion die Distanzmatrix dahingehend modifi-

¹ LANCE, G.N., WILLIAMS, W.T.: A General Theory of Classificatory Sorting Strategies I. Hierarchical Systems. The Computer Journal 9 (1966), S. 373 - 380.

ziert, daß bei der Vereinigung von A_i und A_j zu

$$A'_i = A_i \cup A_j$$

- die j -te Zeile und Spalte von $\mathbf{D}$ gestrichen werden und
- die Gruppenabstände $\tilde{d}_{il}$, $1 \leq l \leq k$, $l \neq i, j$, in der i -ten Zeile und Spalte von $\mathbf{D}$ durch die für A'_i geltenden Abstände $\tilde{d}'_{il}$ ersetzt werden. Diese ergeben sich nach LANCE & WILLIAMS aus

$$\tilde{d}'_{il} = \alpha_i \tilde{d}_{il} + \alpha_j \tilde{d}_{jl} + \beta \tilde{d}_{ij} + \gamma |\tilde{d}_{il} - \tilde{d}_{jl}|, \quad (2.31)$$

wobei die Konstanten α_i , $1 \leq i \leq k$, β und γ vom Verfahren abhängig sind (s. Tab. 2.1).

Durch Variation dieser Parameter kann aus der Formel (2.31) über die o.a. Methoden hinaus gleich eine ganze Schar von weiteren Verfahren konstruiert werden.

- (8) Für die flexible Strategie wird $\alpha_i = \alpha_j = \alpha$, $\gamma = 0$ und $2\alpha + \beta = 1$ gesetzt, wobei α zweckmäßigerweise zwischen 0,5 und 1,0 gewählt wird¹.

Da bei allen Verfahren immer diejenigen Cluster A_i und A_j vereinigt werden, die sich am wenigsten unterscheiden, für die also $\tilde{d}_{ij}$ minimal ist, ist gleichzeitig $A'_i = A_i \cup A_j$ auch das Cluster mit der größten Heterogenität in der aktuellen Partition. $\tilde{d}_{ij}$ ist daher nicht nur ein Maß für die Heterogenität des Clusters A'_i , sondern kann auch als Heterogenitätsmaß für die ganze Partition verwendet werden, da es die maximale Heterogenität der Cluster dieser Partition angibt.

1 In STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 81 wird die Wirkung der Parametervariation etwas genauer beschrieben, gute Erfahrungen wurden mit Werten von $0,6 \leq \alpha \leq 0,7$ gemacht

Verfahren	α_i	α_j	Parameter β	γ	Neue Distanzen $\tilde{d}_{il}'$
(1) Single Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$\min\{\tilde{d}_{il}, \tilde{d}_{jl}\}$
(2) Complete Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\max\{\tilde{d}_{il}, \tilde{d}_{jl}\}$
(3) Average Linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{2}(\tilde{d}_{il} + \tilde{d}_{jl})$
(4) Weighted Average Linkage	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0	$\frac{1}{n_i + n_j}(n_i \tilde{d}_{il} + n_j \tilde{d}_{jl})$
(5) Median-Verf.	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0	$\frac{1}{2}(\tilde{d}_{il} + \tilde{d}_{jl}) - \frac{1}{4}\tilde{d}_{ij}$
(6) Centroid-Verf.	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$\frac{-n_i n_j}{(n_i + n_j)^2}$	0	$\frac{1}{n_i + n_j}(n_i \tilde{d}_{il} + n_j \tilde{d}_{jl}) - \frac{n_i n_j}{(n_i + n_j)^2} \tilde{d}_{ij}$
(7) Wards Verf.	$\frac{n_i + n_l}{n_i + n_j + n_l}$	$\frac{n_j + n_l}{n_i + n_j + n_l}$	$\frac{-n_l}{n_i + n_j + n_l}$	0	$\frac{1}{n_i + n_j + n_l} \left[(n_i + n_l) \tilde{d}_{il} + (n_j + n_l) \tilde{d}_{jl} - n_l \tilde{d}_{ij} \right]$
(8) Flexible Strategie	α	α	$1 - 2\alpha$	0	$\alpha(\tilde{d}_{il} + \tilde{d}_{jl}) + (1 - 2\alpha)\tilde{d}_{ij}$

Tab 2.1: Parameter von 8 hierarchischen Verfahren nach der Formel von LANCE&WILLIAMS¹

1 In Anlehnung an STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 77

Durch die Benutzung der Rekursionsformel (2.31) ist es möglich, alle acht Verfahren mit demselben Algorithmus durchzuführen, indem lediglich bei der Aktualisierung der Clusterabstände $\tilde{d}_{ij}$ die speziellen Werte der Parameter α_i , α_j , β und γ benutzt werden. Für Zwecke dieser Arbeit wurde das von STEINHAUSEN / LANGER implementierte Unterprogramm AGGLOM¹ benutzt, das nach diesem auch von mehreren anderen Autoren in ähnlichen Programmen verwendeten Prinzip arbeitet.

1 In STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 83 f. Vgl. auch z.B. SPÄTH, H.: Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion. a.a.O., Subroutine HIERCL auf S. 172 f.

2.3 ITERATIVE VERFAHREN

Die zweite große Gruppe von Clusteranalyseverfahren bilden diejenigen, die eine vorgegebene Zielfunktion optimieren. Genauso vielfältig wie die Vorstellungen über die "Nähe" von Clustern, die in den agglomerativen hierarchischen Verfahren benutzt werden, sind die Meinungen darüber, was die Güte einer Partition ausmacht und wie diese Güte mit einem Optimalitätskriterium, d.h. durch eine einzige, aus der Datenmatrix X und der Partition α zu berechnende Zahl ausgedrückt werden kann. Einige der gebräuchlichen Optimalitätskriterien können durch mathematisch-statistische Ansätze konstruiert werden; andere dagegen werden ganz oder teilweise¹ von plausiblen Überlegungen geprägt.

Zum Beispiel formuliert BOCK² die Aufgabe, eine vorhandene "richtige Struktur" zu erkennen, in einem stochastischen Modell: Darin wird unterstellt, daß die n beobachteten Objekte lediglich eine Stichprobe aus einer größeren Grundgesamtheit darstellen; d.h. die Beobachtungsvektoren x_i , $1 \leq i \leq n$, sind Realisationen von stochastisch-unabhängigen, m -dimensionalen Zufallsvariablen $\mathfrak{X}_i$, $1 \leq i \leq n$, mit einer bestimmten Verteilung. Das Problem besteht nun darin, diejenigen Zufallsvariablen herauszufinden, die dieselbe Verteilung besitzen. Unter der Voraussetzung, daß alle $\mathfrak{X}_i$ einer m -dimensionalen Normalverteilung mit Mittelwertvektoren $\mu_i \in \mathbb{R}^m$ und Kovarianzmatrizen $\Sigma_i \in \mathbb{R}^{(m,m)}$ unterliegen, also

$$\mathfrak{X}_i \sim N_m(\mu_i, \Sigma_i) \quad , \quad 1 \leq i \leq n, \quad (2.32)$$

und die wahre Anzahl k , $k \in \mathbb{N}$, von unterschiedlichen Verteilungen bekannt ist, können nun aus der gemeinsamen

1 s. z.B. (2.59) auf S. 42

2 BOCK, H.H.: Automatische Klassifikation. a.a.O., S. 113 ff.

Verteilung aller $\mathbf{x}_i$ Maximum-Likelihood-Verfahren für die Schätzung der richtigen Partition $\mathbf{A}$ abgeleitet werden.

- (1) Wenn angenommen wird, daß alle Kovarianzmatrizen identisch sind und einem Vielfachen der m -reihigen Einheitsmatrix $\mathbf{I}_m$ entsprechen, wenn also für alle Cluster A_j

$$\mathbf{x}_i \sim N_m(\boldsymbol{\mu}_j, \sigma^2 \mathbf{I}_m) \quad \forall i \in A_j, 1 \leq j \leq k, \quad (2.33)$$

mit $\sigma^2 \in \mathbb{R}^+$ gilt, ist diejenige Partition eine Maximum-Likelihood-Schätzung, für die die als Varianzkriterium bezeichnete Größe

$$K_1(\mathbf{A}, \mathbf{X}) = \sum_{j=1}^k \sum_{i \in A_j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2 \rightarrow \min \quad (2.34)$$

minimal wird. In dem hier unterstellten Fall, daß die m Variablen unkorreliert sind und die gleiche Varianz aufweisen, d.h. bildlich, daß die Cluster als gleichgroße Kugeln im Merkmalsraum erscheinen, müßte ein Maximum-Likelihood-Verfahren also bestrebt sein, die Cluster so zu bilden, daß $K_1(\mathbf{A}, \mathbf{X})$ dabei einen möglichst geringen Wert annimmt.

- (2) Wird zugelassen, daß die Kugeln ungleich groß sind, wobei die m Variablen aber weiterhin unkorreliert sind, also

$$\mathbf{x}_i \sim N_m(\boldsymbol{\mu}_j, \sigma_j^2 \mathbf{I}_m) \quad \forall i \in A_j, 1 \leq j \leq k, \quad (2.35)$$

und $\sigma_j^2 \in \mathbb{R}^+, 1 \leq j \leq k$, so minimiert eine Maximum-Likelihood-Schätzung

$$K_2(\mathbf{A}, \mathbf{X}) = \prod_{j=1}^k \left[\frac{1}{n_j} \sum_{i \in A_j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2 \right]^{n_j} \rightarrow \min. \quad (2.36)$$

Dieses Kriterium soll hier Kriterium der Gewichteten Varianzen genannt werden.

Sobald die Annahme der kugelförmigen Cluster, die immer bei stochastisch unabhängigen Variablen gegeben wäre, fallen gelassen wird, sind für die Schätzung der richtigen Partition plausiblerweise nicht nur die in $K_1(\mathbf{0}, \mathbf{X})$ und $K_2(\mathbf{0}, \mathbf{X})$ enthaltenen Varianzen der Variablen, sondern auch alle Kovarianzen von Bedeutung. Diese können aus den folgenden sog. Streuungsmatrizen (scatter matrices) geschätzt werden:

$$\mathbf{T} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^t (\mathbf{x}_i - \bar{\mathbf{x}}) \in \mathbf{R}^{(m,m)} \quad (2.37)$$

$$\mathbf{W} = \sum_{j=1}^k \sum_{i \in A_j} (\mathbf{x}_i - \bar{\mathbf{x}}_j)^t (\mathbf{x}_i - \bar{\mathbf{x}}_j) \in \mathbf{R}^{(m,m)} \quad (2.38)$$

$$\mathbf{B} = \sum_{j=1}^k n_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^t (\bar{\mathbf{x}}_j - \bar{\mathbf{x}}) \in \mathbf{R}^{(m,m)} \quad (2.39)$$

Dabei ist

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad (2.40)$$

der Mittelwertvektor der gesamten durch $\mathbf{x}_1, \dots, \mathbf{x}_n$ beschriebenen Objektmenge. Von diesen Matrizen enthält nun

$\mathbf{T}$ (= total scatter matrix) die Gesamtstreuung des Datensatzes ohne Berücksichtigung einer Gruppeneinteilung;

$\mathbf{W}$ (= within scatter matrix) die Streuung der Objekte innerhalb der Gruppen, die sich aus je einer Streuungsmatrix

$$\mathbf{W}_j = \sum_{i \in A_j} (\mathbf{x}_i - \bar{\mathbf{x}}_j)^t (\mathbf{x}_i - \bar{\mathbf{x}}_j) \quad , \quad 1 \leq j \leq k, \quad (2.41)$$

für jedes Cluster A_j , $1 \leq j \leq k$, als

$$\mathbf{W} = \sum_{j=1}^k \mathbf{W}_j \quad (2.42)$$

zusammensetzt, und

$\mathbf{B}$ (= between groups scatter matrix) die Streuung der Gruppen oder genauer der Gruppenmittelwerte $\bar{\mathbf{x}}_j$ untereinander.

Mit Hilfe der so definierten Streuungsmatrizen, zwischen denen die leicht zu zeigende Beziehung

$$\mathbf{T} = \mathbf{W} + \mathbf{B} \quad (2.43)$$

besteht, können die bereits genannten Kriterien $K_1(\mathbf{Q}, \mathbf{X})$ und $K_2(\mathbf{Q}, \mathbf{X})$ nun auch als

$$K_1(\mathbf{Q}, \mathbf{X}) = \text{Spur}(\mathbf{W}) \rightarrow \min \quad (2.44)$$

bzw. äquivalent dazu

$$K_1'(\mathbf{Q}, \mathbf{X}) = \text{Spur}(\mathbf{B}) \rightarrow \max \quad (2.45)$$

und

$$K_2(\mathbf{Q}, \mathbf{X}) = \prod_{j=1}^k \left[\frac{1}{n_j} \text{Spur}(\mathbf{W}_j) \right]^{n_j} \rightarrow \min \quad (2.46)$$

formuliert und berechnet werden.

(3) Läßt man in dem oben beschriebenen Modell eine beliebige, aber für alle Cluster gleiche (positiv definite) Kovarianzmatrix $\Sigma \in \mathbb{R}^{(m,m)}$ zu, sodaß

$$\mathbf{x}_i \sim N_m(\boldsymbol{\mu}_j, \Sigma) \quad \forall i \in A_j, \quad 1 \leq j \leq k \quad (2.47)$$

gilt, so folgt, daß

$$K_3(\alpha, X) = |W| \rightarrow \min, \quad (2.48)$$

das sogenannte Determinantenkriterium, ein Maximum-Likelihood-Kriterium ist.

- (4) Müssen auch noch unterschiedliche Kovarianzenmatrizen $\Sigma_j \in R^{(m,m)}$, $1 \leq j \leq k$ unterstellt werden, also

$$x_i \sim N_m(\mu_j, \Sigma_j) \quad \forall i \in A_j, \quad 1 \leq j \leq k, \quad (2.49)$$

dann muß für eine Maximum-Likelihood-Schätzung

$$K_4(\alpha, X) = \sum_{j=1}^k \left| \frac{1}{n_j} W_j \right|^{n_j} \rightarrow \min \quad (2.50)$$

minimiert werden; dieses Kriterium soll hier als Kriterium der Gewichteten Determinanten bezeichnet werden.

Außer den angegebenen Zielfunktionen für Maximum-Likelihood-Verfahren werden bei BOCK¹ auch noch Zielfunktionen für Bayes-Verfahren abgeleitet, die aber hier nicht weiter berücksichtigt werden sollen, weil sie mit bestimmten Verlustfunktionen und a-Priori-Verteilungen weitaus stärkere Voraussetzungen benötigen und wegen ihrer teilweise äußerst komplizierten Gestalt kaum noch zu interpretieren sind.

Eine andere Möglichkeit, Optimalitätskriterien für die Clusteranalyse abzuleiten, bieten die Teststatistiken, die zur Prüfung linearer Hypothesen verwendet werden. Dabei wird wie oben unterstellt, daß die Beobachtungsvektoren x_i , $1 \leq i \leq n$, Realisationen von stochastisch unabhängigen Zufallsvariablen x_i , $1 \leq i \leq n$, sind, die alle m-dimensional

1 BOCK, H.H.: Automatische Klassifikation. a.a.O., S. 113 ff.

normalverteilt sind mit derselben Kovarianzmatrix $\Sigma \in \mathbb{R}^{(m,m)}$. Darüberhinaus haben jeweils einige der Verteilungen auch den gleichen Mittelwertvektor, d.h. es ist eine Partition $\mathcal{A} = \{A_1, \dots, A_k\}$ mit $k \leq n$ bekannt, für die

$$\mathbf{x}_i \sim N_m(\boldsymbol{\mu}_j, \Sigma) \quad \forall i \in A_j, 1 \leq j \leq k, \quad (2.51)$$

mit $\boldsymbol{\mu}_j \in \mathbb{R}^m$, $1 \leq j \leq k$, gilt. Mit der Varianzanalyse kann nun unter anderem die Hypothese untersucht werden, daß die Mittelwertvektoren $\boldsymbol{\mu}_j$, $1 \leq j \leq k$, gar nicht verschieden sind, wonach dann alle $\mathbf{x}_i$, $1 \leq i \leq n$, dieselbe Verteilung hätten und die Partition $\mathcal{A}$ ein reines Artefakt wäre. Für den varianzanalytischen Test der Hypothese

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_k \quad (2.52)$$

wird aus den beobachteten Vektoren $\mathbf{x}_i$, $1 \leq i \leq n$, eine Prüfgröße, d.h. eine reellwertige Funktion, berechnet, die kleine Werte annimmt, wenn die Beobachtungen für H_0 sprechen. Für das Clusteranalyseproblem, bei dem die Partition $\mathcal{A}$ erst gefunden werden muß, ist es von daher naheliegend, die Cluster so zu bilden, daß diese Prüfgröße möglichst groß wird, weil man dadurch einen im Sinne des Tests der Hypothese (2.52) größtmöglichen Unterschied zwischen den Clustern erzeugt.

Im univariaten Fall gibt es für das o.a. Testproblem

einen gleichmäßig besten invarianten¹ Test, der die Prüfgröße

$$\tilde{f} = \frac{\frac{1}{k-1} \sum_{j=1}^k \|\bar{\mathbf{x}}_j - \bar{\mathbf{x}}\|^2}{\frac{1}{n-k} \sum_{j=1}^k \sum_{i \in A_j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2} \quad (2.53)$$

benutzt, die in Matrizenschreibweise² auch als

$$\tilde{f} = \frac{n-k}{k-1} \frac{\mathbf{B}}{\mathbf{W}} = \frac{n-k}{k-1} \mathbf{W}^{-1} \mathbf{B} \quad (2.54)$$

geschrieben werden könnte³. Für eine Clusteranalyse müßte nun $\tilde{f}$ maximiert werden. Im multivariaten Fall kann leider die Menge der möglichen Tests auch durch Invarianzforderungen nicht so weit eingeschränkt werden, daß unter den verbleibenden Möglichkeiten ein gleichmäßig bester Test gefunden werden könnte. Es kann aber gezeigt werden, daß alle invarianten Teststatistiken⁴ nur von den Wurzeln $\lambda_1, \dots, \lambda_m$ der Determinantengleichung

$$|\mathbf{B} - \lambda \mathbf{W}| = 0 \quad (2.55)$$

abhängen, die gleichzeitig auch die Eigenwerte von $\mathbf{W}^{-1} \mathbf{B}$ sind.

-
- 1 Dieser Test ist invariant gegenüber Transformationen der Form $\mathbf{f}(\mathbf{X}) = \mathbf{c} \mathbf{X} \mathbf{L} + \mathbf{d}$ für $\mathbf{X} \in \mathbb{R}^m$, mit $\mathbf{c} \in \mathbb{R} - \{0\}$, $\mathbf{d} \in \mathbb{R}^m$ und einer orthogonalen Matrix $\mathbf{L} \in \mathbb{R}^{(m,m)}$. Vgl. z.B. WITTING, H., NÖLLE, G.: Angewandte Mathematische Statistik. Stuttgart 1970, Satz 1.5 auf S.37 und Beispiel 1.10 auf S.32. Zur einfachen Zerlegung siehe Beispiel 1.4 auf S.19.
 - 2 Obwohl die Vektor- und Matrizenschreibweise im univariaten Fall nicht nötig ist, wird sie hier beibehalten, um die Zusammenhänge mit dem multivariaten Fall besser sichtbar zu machen.
 - 3 Der Quotient zweier Matrizen ist natürlich nur im univariaten Fall definiert.
 - 4 Vgl. die Herleitung von Teststatistiken zur Prüfung linearer Hypothesen z.B. in AHRENS, H., LÄUTER, J.: Mehrdimensionale Varianzanalyse, Hypothesenprüfung, Dimensionserniedrigung, Diskrimination. Berlin 1974, S. 33 ff.

Da ein gleichmäßig bester Test nicht existiert, sind verschiedene Funktionen der Eigenwerte $\lambda_1, \dots, \lambda_m$ als Prüfgrößen vorgeschlagen worden, die auch als Optimalitätskriterien bei Clusteranalysen Verwendung finden können. Die folgenden drei Kriterien, die als die gebräuchlichsten hier berücksichtigt werden sollen, sind im univariaten Fall zur Prüfgröße (2.53) bzw. (2.54) äquivalent. Für $m > 1$ ergeben sich jedoch echt unterschiedliche Tests¹ und bei Verwendung als Zielfunktion bei Clusteranalysen auch verschiedene Ergebnisse.

(3') Von WILKS² wurde das sog. WILKS'sche Lambda

$$K'_3(\mathbf{A}, \mathbf{X}) = \sum_{j=1}^m \frac{1}{1 + \lambda_j} = \frac{|\mathbf{W}|}{|\mathbf{T}|} \rightarrow \min \quad (2.56)$$

vorgeschlagen, das zu dem bereits eingeführten Determinantenkriterium $K_3(\mathbf{A}, \mathbf{X})$ äquivalent ist, weil $\mathbf{T}$ und damit auch $|\mathbf{T}|$ nicht von der Partition abhängen.

(5) HOTELLINGS-T₀² stellt eine Verallgemeinerung der Prüfgröße des univariaten F-Tests dar und hat die folgende Form³:

$$K_5(\mathbf{A}, \mathbf{X}) = \sum_{j=1}^m \lambda_j = \text{Spur}(\mathbf{W}^{-1}\mathbf{T}) \rightarrow \max. \quad (2.57)$$

Dieses Kriterium wird auch als Spurkriterium bezeichnet.

(6) Gelegentlich wird auch der Größte Eigenwert als Kriterium vorgeschlagen⁴, d.h.

$$K_6(\mathbf{A}, \mathbf{X}) = \max\{\lambda_j / 1 \leq j \leq m\} \rightarrow \max \quad (2.58)$$

1 S. AHRENS, H., LÄUTER, J.: Mehrdimensionale Varianzanalyse, Hypothesenprüfung, Dimensionserniedrigung, Diskrimination. a.a.O., S. 38

2 WILKS, S.S.: Certain Generalizations in the Analysis of Variance. Biometrika 24 (1932), S. 471-494.

3 HOTELLING, A.: A Generalized T-test and a Measure of Multivariate Dispersion. In: Proceeding of the Second Berkeley Symposium of Mathematical Statistics and Probability, 1959, S. 23-41.

4 S.z.B. ROY, S.N.: Some Aspects of Multivariate Analysis. New York 1957. Dort wird eine ganze Reihe von Tests mit Hilfe der Verteilung des größten Eigenwertes konstruiert.

Darüberhinaus kann selbstverständlich eine Reihe weiterer Kriterien aus den Eigenwerten von $\mathbf{W}^{-1}\mathbf{B}$ gebildet werden, von denen aber einige ohnehin zu einem der bereits genannten Kriterien äquivalent sind und die anderen nur selten genannt werden.

Viele der gebräuchlichen Zielfunktionen für Clusteranalysen sind weniger mit Hilfe mathematischer Modelle zustande gekommen als mit größtenteils heuristischen Überlegungen oder aus einer Mischung von plausiblen Argumenten und theoretischen Kenntnissen.

- (7) Zum Beispiel ist es naheliegend, ein Cluster A_j dann als homogen anzusehen, wenn der Abstand der einzelnen Objekte zu einem Gruppenrepräsentanten $\mathbf{x}_j^*$ möglichst gering wird. Daraus resultiert eine zu minimierende Zielfunktion der Form

$$\sum_{j=1}^k \sum_{i \in A_j} d(\mathbf{x}_i, \mathbf{x}_j^*) \quad . \quad (2.59)$$

Dieses Kriterium findet insbesondere dann Verwendung, wenn d zur Gruppe der sog. L_p -Metriken gehört, bei denen der Abstand $d^p(\mathbf{x}, \mathbf{y})$ zwischen zwei Vektoren $\mathbf{x} = (x_1, \dots, x_m)$ und $\mathbf{y} = (y_1, \dots, y_m)$ als

$$d^p(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^m |x_j - y_j|^p \right]^{1/p} \quad \text{mit } p \in \mathbb{R}^+ \quad (2.60)$$

definiert wird; als Spezialfälle sind in dieser Gruppe z.B. mit $p = 2$ die in (2.22) bereits genannte euklidische Distanz und mit $p = 1$ die sog. City-Block-Distanz enthalten. Als Beispiel einer solchen Zielfunktion soll hier

$$K_7(\mathbf{A}, \mathbf{X}) = \sum_{j=1}^k \sum_{i \in A_j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\| \rightarrow \min \quad (2.61)$$

mit einbezogen werden; in diesem Kriterium wird die Homogenität der Cluster durch die euklidische Distanz der Objekte zum jeweiligen Clustermittelwert (Centroid)

gemessen¹. Es soll hier als Abstandssummenkriterium bezeichnet werden.

- (8) Relativ neu ist ein Vorschlag², der wie bei der Ableitung von

$$K_4(\alpha, X) = \prod_{j=1}^k \left| \frac{1}{n_j} \mathbf{w}_j \right|^{n_j} \rightarrow \min \quad (2.62)$$

davon ausgeht, daß die einzelnen Cluster unterschiedlich ausgerichtete Achsen aufweisen. Daher ist für die Clusteranalyse nicht so sehr die gesamte Innergruppenstreuungsmatrix $\mathbf{W}$ relevant, wie sie z.B. in

$$K_3(\alpha, X) = |\mathbf{W}| = \left| \sum_{j=1}^k \mathbf{w}_j \right| \rightarrow \min \quad (2.63)$$

benutzt wird, sondern es ist mehr Gewicht auf die Verschiedenartigkeit der Streuungsmatrizen $\mathbf{W}_j$ der einzelnen Cluster A_j zu legen. Abweichend von dem entscheidungstheoretisch motivierten Kriterium (2.62) lautet jedoch das Kriterium der adaptiven Distanzen

$$K_8(\alpha, X) = \sum_{j=1}^k |\mathbf{w}_j|^{1/m} \rightarrow \min \quad . \quad (2.64)$$

Für die hier genannten Zielfunktionen $K_1(\alpha, X)$ bis $K_8(\alpha, X)$ sind - bis auf eine Ausnahme mit sehr beschränkter Anwendungsmöglichkeit - keine Algorithmen bekannt, die mit Sicherheit ein globales Optimum finden. Zwar stünde grundsätzlich als Notlösung die vollständige Enumeration sämtlicher Möglichkeiten im Hintergrund, in Anbetracht der riesigen Anzahl von Partitionen schon bei kleineren Objektmengen scheidet dieses Verfahren jedoch selbst bei Benutzung modernster Rechenanlagen aus. Bezeichnet man nämlich mit

1 Für $p=1$ wird häufig auch vorgeschlagen, den Median als Clustermittelpunkt einzusetzen.

2 SPÄTH, H., MÜLLER, R.: Das Austauschverfahren für die skalierungsinvariante Methode der adaptiven Distanzen in der Cluster-Analyse. In SPÄTH, H. (Hrsg.): Ausgewählte Operations Research Software in FORTRAN. München, Wien 1979, S. 143-163.

- $S(n,k)$ die Anzahl der Partitionen der Länge k bei einer Menge mit n Objekten (STIRLING'sche Zahlen zweiter Art) und mit
 - $B(n)$ die Anzahl der Partitionen beliebiger Länge bei n Objekten (BELL'sche Zahlen),
- so gilt

$$S(n,k) = \frac{1}{k!} \sum_{i=0}^k (-1)^i \binom{k}{i} (k-i)^n, \quad (2.65)$$

$$B(n) = \sum_{i=1}^n S(n,i) \quad . \quad (2.66)$$

Einen Eindruck von der Größenordnung dieser Zahlen geben die Tabellen 2.2 und 2.3.

Es ist leicht zu errechnen, daß selbst bei kleineren Objektzahlen und schnellsten Rechenanlagen die vollständige Enumeration Jahre dauern würde. Einen Kompromiß zwischen der Forderung nach optimalen Lösungen und der Durchführbarkeit stellen die iterativen Verfahren dar, die durch fortgesetzte Umordnung von Objekten eine Zielfunktion verbessern. Zwar bleiben sie unter Umständen an einem lokalen Optimum hängen, das sie durch Umgruppierung eines einzelnen Elementes nicht mehr verlassen können, dafür sind sie so schnell, daß es vertretbar ist, mehrere solche suboptimale Lösungen zu erzeugen, sodaß man sicher sein kann, zumindest eine dem Optimum relativ nahe Lösung dabei zu haben.

$\begin{smallmatrix} n \\ k \end{smallmatrix}$	10	20	25
1	1	1	1
2	511	524 287	16 777 215
3	9 330	580 606 446	141 197 991 025
4	34 105	45 232 115 901	46 771 289 738 810
5	42 525	749 206 090 500	2 436 684 974 110 751
6	22 827	4 306 078 895 384	37 026 417 000 002 430
7	5 880	11 143 554 045 652	227 832 482 998 716 310
8	750	15 170 932 662 679	690 223 721 118 368 580
9	45	12 011 282 644 725	1 167 921 451 092 973 005
10	1	5 917 584 964 655	1 203 163 392 175 387 500
.	.	.	.
.	.	.	.
15	.	452 329 200	4 299 394 655 347 200
.	.	.	.
.	.	.	.
20	1	.	6 220 194 750
.	.	.	.
.	.	.	.
25	.	.	1

Tab 2.2: STIRLING'sche Zahlen zweiter Art¹

n	B(n)
1	1
2	2
3	5
4	15
5	52
6	203
7	877
8	4 140
9	21 147
10	115 957
.	.
.	.
15	1 382 958 545
.	.
.	.
20	51 724 158 235 372
.	.
.	.
25	4 638 590 322 228 999 353

Tab 2.3: BELL'sche Zahlen

1 Auszug aus: ABRAMOWITZ, M., STEGUN, I.A. (Hrsg.): Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. National Bureau of Standards. Applied Mathematics Series Bd. 55, 10. Korr. Aufl., Washington 1972, S. 835

Die Minimaldistanzverfahren verbessern eine gegebene Partition, ausgehend von einer oft zufälligen, häufig auch von einem hierarchischen Verfahren bestimmten Startpartition, durch Wiederholung der folgenden Schritte:

- Berechnen aller Gruppenmittelpunkte und
- Umgruppierung jedes Elementes in das Cluster mit dem nächstgelegenen Mittelpunkt.

Das Verfahren ist beendet, wenn in einem Durchgang kein Element mehr umgeordnet wurde. Offensichtlich wird die zu optimierende Zielfunktion nur indirekt durch die Art und Weise der Bestimmung von Gruppenmittelpunkten und das Distanzmaß zwischen Element und Mittelpunkt in diesem Verfahren berücksichtigt. Daher ist auch nicht immer unmittelbar zu erkennen, welche Zielfunktion ein Minimaldistanzverfahren im Einzelfall optimiert.

Eine einfache Rechnung¹ zeigt jedoch, daß z.B. das Varianzkriterium $K_1(\mathcal{O}, \mathbf{X})$ bei Umordnungen nach dem Minimaldistanzprinzip zumindest nicht verschlechtert werden kann, wenn als Gruppenmittelpunkte die Centroide $\bar{\mathbf{x}}_j$, $1 \leq j \leq k$ und als Distanzmaß die euklidische Distanz (2.29) gewählt werden, d.h. ein solcher Algorithmus erreicht zumindest ein Suboptimum des Varianzkriteriums. Zum Vergleich mit anderen Verfahren wurde hier das von STEINHAUSEN/LANGER implementierte Unterprogramm MINDS1 benutzt².

Im Gegensatz zu den Minimaldistanzverfahren orientieren sich die Austauschverfahren direkt an der Zielfunktion und sind daher zur Verbesserung beliebiger Optimalitätskriterien geeignet. Sie überprüfen einfach Objekt für Objekt, ob durch die Umgruppierung in ein anderes Cluster eine Verbesserung der Zielfunktion erreicht wird und verschieben das Objekt gegebenenfalls in das Cluster mit der

1 STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 107

2 ebenda, S. 109

größten Verbesserung. Zur Minimierung des Varianzkriteriums haben STEINHAUSEN und LANGER¹ bzw. STEINHAUSEN² verschiedene Versionen eines Austauschalgorithmus programmiert, von denen hier die als HLCMB1 bezeichnete Version verwendet wurde. Da für einige der übrigen 7 genannten Kriterien jedoch keine optimierenden Algorithmen zur Verfügung stehen, wurde weiterhin vom Verfasser ein Austauschalgorithmus programmiert, der wahlweise alle acht Zielfunktionen zur Steuerung der Clusteranalyse verwenden kann (s. den Programmablaufplan in Abb. 2.3 und 2.4).

Das Programm erhebt keinerlei Anspruch auf Eleganz oder Schnelligkeit; bei Verfahren, die auf spezielle Kriterien ausgerichtet sind, können die Austauschschritte und die jeweils erforderliche Neuberechnung des Kriteriums u.U. erheblich verkürzt und beschleunigt werden. Jedoch kann der Algorithmus in der vorliegenden Form praktisch jede erdenkliche Zielfunktion verarbeiten, ohne daß er dazu wesentlich verändert werden müßte; daher kann für alle o.g. Zielfunktionen dasselbe Programm verwendet werden, indem lediglich bei der Berechnung der Zielfunktionswerte ein entsprechender Abschnitt angesprungen wird.

Für die Minimierung des Varianzkriteriums, das wohl ohnehin am häufigsten in der Clusteranalyse verwendet wird, existiert weiterhin ein Branch-and-Bound-Verfahren, das mit Sicherheit ein globales Minimum findet. Für kleinere Objekt- und Clusterzahlen, d.h. $n \leq 25$ und $k \leq 6$ konnte hier ein von HERDEN/STEINHAUSEN implementiertes Programm benutzt werden, dessen Hauptbestandteil die Subroutine

1 STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 119 f.

2 STEINHAUSEN, D.: Clusteranalyse großer Datenmengen - effektive Verfahren und Programme zur Clusterung quantitativer Daten. Schriftenreihe des Rechenzentrums der Universität Münster Nr. 18, 1977.

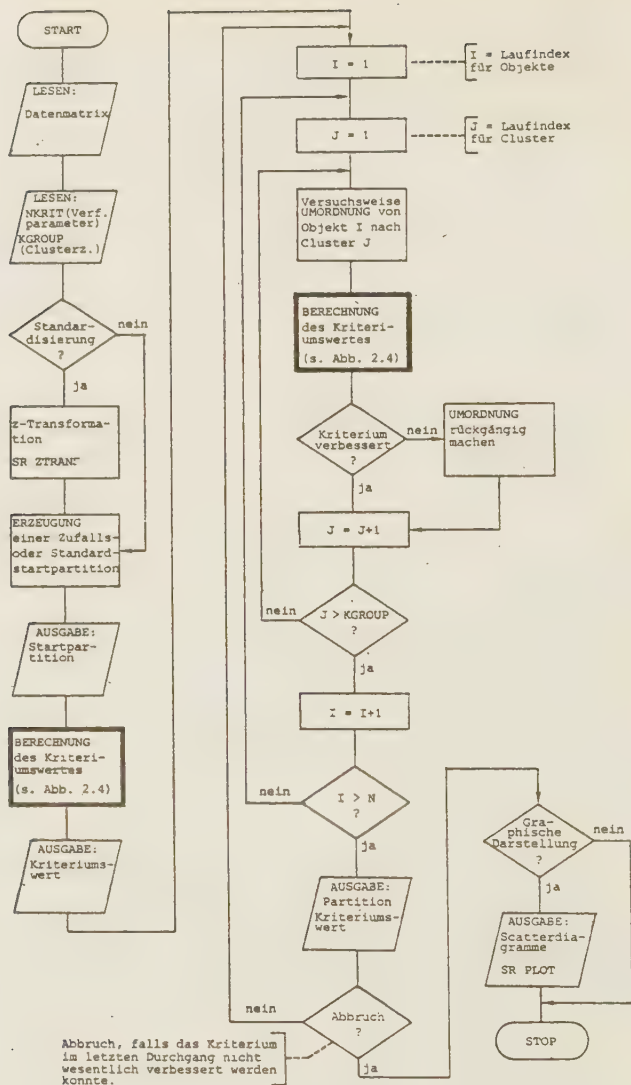


Abb 2.3: Programmablaufplan eines Austauschalgorithmus, der zur Optimierung von acht verschiedenen Optimalitätskriterien geeignet ist. Der Teil zur Berechnung des Kriteriums-wertes ist in Abb. 2.4 gesondert dargestellt.

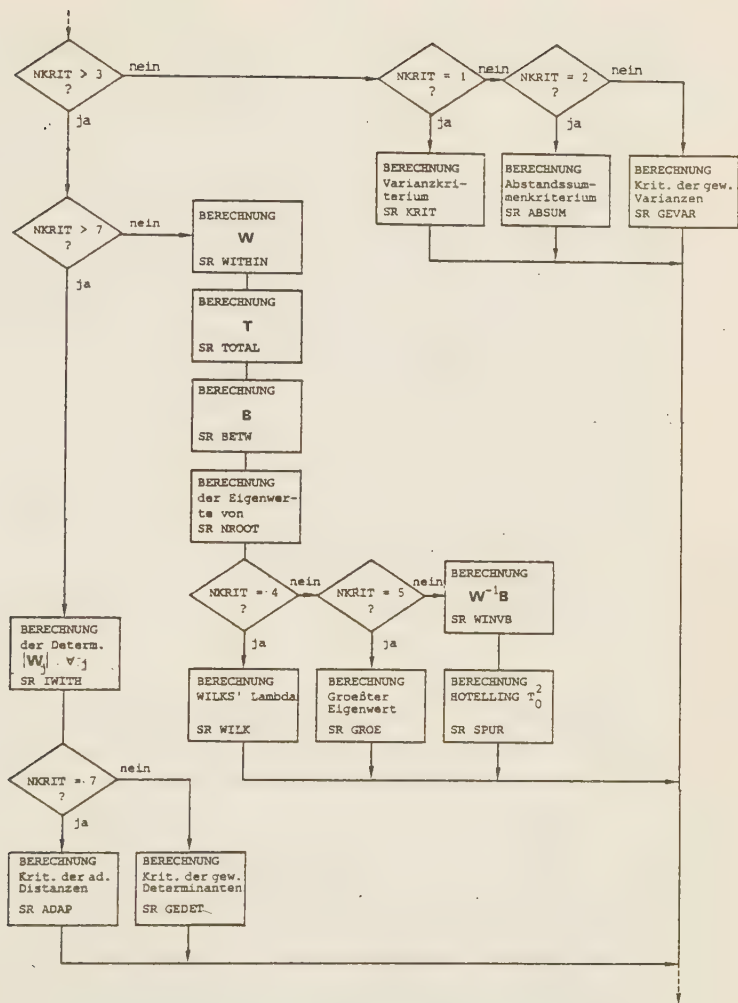


Abb 2.4: Teil des Programmablaufplans in Abb. 2.3 zur Berechnung des Kriteriumswertes¹

1 Im Programm werden neben den selbst erstellten Unterprogrammen die Subroutines RANDU zur Erzeugung gleichverteilter Zufallszahlen, NROOT zur Berechnung von Eigenwerten und (innerhalb des Unterprogramms WINVB zur Berechnung von $W^{-1}B$) MINV zur Matrizeninversion aus dem Programmpaket SSP aufgerufen. Vgl. SCIENTIFIC SUBROUTINE PACKAGE (SSP), Version III, Programmer's Manual, o.O., 5. Aufl 1970, S. 77 (RANDU), S. 166 (NROOT) und S. 118 (MINV). Ferner wird zur Standardisierung der Datenmatrix die Subroutine ZTRANS aus der Programmsammlung von STEINHAUSEN/LANGER zur Clusteranalyse aufgerufen. S. STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 88.

BBOUND¹ ist. Leider steigt die Bearbeitungsdauer mit den Objekt- und Clusterzahlen stark an² und hängt darüberhinaus auch davon ab, wie deutlich die zu clusternde Objektmenge in Teile zerfällt. Je homogener nämlich die Objektmenge ist, um so schwerer fällt es dem Entscheidungsbaumverfahren, Äste zu sperren, die nicht zum Optimum führen können; im Extremfall führt dies zu einer vollständigen Enumeration sämtlicher Lösungen. Daher werden auch für das Varianzkriterium bei realistischen Größenordnungen der Objektzahlen weiterhin Minimaldistanz- und Austauschverfahren verwendet werden müssen.

1 vgl. HERDEN, W., STEINHAUSEN, D.: Ein Verfahren zur Berechnung eines globalen Minimums beim Varianzkriterium in der Clusteranalyse. In SPÄTH, H. (Hrsg.): Ausgewählte Operations Research Software in FORTRAN. München 1979. S. 115-142.

2 vgl. STEINHAUSEN, D.: Ein global optimaler Clusteranalysealgorithmus zur Optimierung von Zielfunktionen. Proceedings in Operations Research 7, Vorträge der Jahrestagung 1977 DGOR, Würzburg, Wien 1978, S. 97-104.

3 PROBLEME DER CLUSTERANALYSE

3.1 AUSWAHL UND AUFBEREITUNG DER DATEN

Für die sinnvolle Anwendung numerischer Klassifikationsverfahren ist es wichtig, daß der Benutzer sich darüber im Klaren ist, daß nicht nur von verschiedenen Verfahren verschiedene Ergebnisse produziert werden, sondern bereits in der Auswahl bzw. der Aufbereitung der Eingabedaten umfangreiche Möglichkeiten liegen, das spätere Endergebnis zu beeinflussen. Für den Routinier bietet sich hier ein Ansatzpunkt, den Ablauf der Clusteranalyse im Sinne seiner Absicht zu steuern, für den unerfahrenen Benutzer, der sich die technischen Möglichkeiten der Clusteranalyse "blind" - d.h. ohne in den theoretischen Hintergrund einzudringen - zunutze machen will, liegt hier eher die Gefahr, unwissentlich ein Ergebnis in einer Weise vorzuprägen, die kein Verfahren mehr ausgleichen kann.

Die Anzahl und Auswahl der Objekte steht in aller Regel vor der Clusteranalyse nicht zur Debatte. Immerhin sollten aber zwei Gesichtspunkte berücksichtigt werden:

- Stellt die zu clusternde Objektmenge nur eine Stichprobe aus einer größeren Gesamtheit dar, wie dies z.B. in der Marktforschung häufig vorkommt, so können nur dann aus der gefundenen Klasseneinteilung begründete Schlüsse über die Grundgesamtheit gezogen werden, wenn jedes einzelne Cluster genügend Elemente enthält, um die entsprechende Teilgesamtheit ausreichend zu repräsentieren. Die Gesamtzahl der Objekte ist daher weniger wichtig als die Mächtigkeit jedes einzelnen Clusters.
- Ein oder mehrere in der Objektmenge enthaltene sog. Ausreißer können die Ergebnisse stark verändern, indem

sie z.B. eigene Cluster bilden oder die räumliche Ausrichtung der Cluster verzerren. Dabei üben ungerechtfertigterweise gerade solche Objekte, die eine sehr unwahrscheinliche Kombination von Merkmalsausprägungen aufweisen, einen besonders starken Einfluß aus. Um dies zu vermeiden ist die Elimination von Ausreißern durch vorgelagerte Ausreißertests¹ zu erwägen. Andererseits sind aber einige Clusteranalyseverfahren wie z.B. Single-Linkage selbst gute Methoden zur Entdeckung von Ausreißern, weil dadurch, daß ein Element erst sehr spät überhaupt in ein Cluster aufgenommen wird, signalisiert wird, daß dieses Element von allen anderen weit entfernt ist (s. Abb. 3.1).

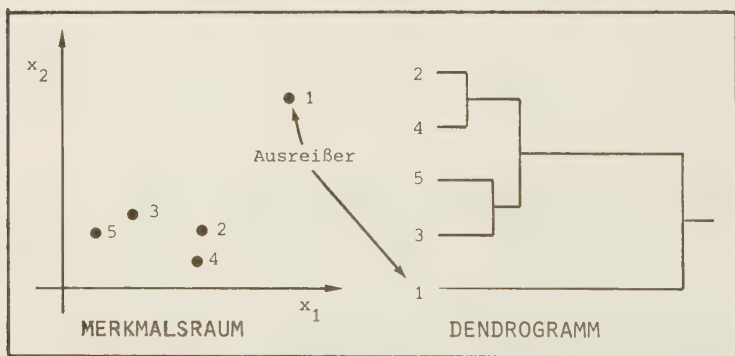


Abb. 3.1.: Erkennung eines Ausreißers im Dendrogramm

¹ s. z.B. ROHLF, J.F.: Generalization of the Gap Test for the Detection of Multivariate Outliers. Biometrics 31 (1975), S. 93-101.

Wesentlich mehr Einflußmöglichkeiten als bei der Auswahl ergeben sich bei der Repräsentation der Objekte durch Merkmalsausprägungen oder Ähnlichkeitsbeziehungen. Bei der Ermittlung einer Datenmatrix **X** zur Eingabe in ein iteratives Verfahren oder auch zur Weiterverarbeitung zu einer Distanzmatrix für ein hierarchisches Verfahren treten mehrere Fragen auf, für die es keine eindeutigen Antworten gibt, die aber trotzdem für das Ergebnis der Clusteranalyse wesentlich sind:

- Welche und wieviele Merkmale sollen für die Klassifikation benutzt werden?
- Wie wirken sich unterschiedliche Skalenniveaus aus?
- Wie müssen Abhängigkeiten zwischen den Merkmalen berücksichtigt werden?
- Wie können die Merkmale angemessen gewichtet werden?

Gelegentlich wird vermutet, daß es - in Analogie zu einem notwendigen Stichprobenumfang in der Entscheidungstheorie - auch eine Mindestzahl von Variablen gibt, die bei einer Clusteranalyse berücksichtigt werden sollten¹. Tatsächlich ist es plausibel anzunehmen, daß bei der Hinzunahme immer weiterer Merkmale irgendwann alle relevanten Merkmale erfaßt und weitere Merkmale redundant oder unwichtig sind und keinen Einfluß mehr auf die Clusteranalyse ausüben. Eine solche Variablenzahl, bei der sich das Ergebnis der Clusteranalyse stabilisiert, wäre jedoch, auch wenn es sie gäbe, vermutlich von einer Objektmenge zur anderen verschieden und nicht a priori festzulegen, sodaß man sich nur darauf beschränken kann, alle als relevant erachteten Variablen zu erfassen. Die

1 vgl. SNEATH, P.H.A., SOKAL, R.R.: Numerical Taxonomy.
a.a.O., S. 106 ff.

Relevanz eines Merkmals wiederum kann mit Sicherheit nicht aus der Theorie der Clusteranalyse, sondern nur aus dem konkreten Anwendungsbezug heraus beurteilt werden. Zumindest können aber einige Merkmalstypen benannt werden, die nach Möglichkeit nicht verwendet werden sollten¹:

- Merkmale, die im Untersuchungszusammenhang keine Bedeutung haben, sind nicht nur keine Hilfe, sondern können genauso starke, aber unerwünschte Einflüsse wie jedes andere Merkmal ausüben.
- Funktional abhängige Merkmale, z.B. Umsatz und Umsatzsteuer, sind redundant und verbessern daher die Informationsbasis nicht. Sie bewirken jedoch eine u.U. verdeckte stärkere Gewichtung derselben Variablen. Auch bei nur teilweisen Abhängigkeiten, etwa zwischen Umsatzsteuer und Gesamtsteuerbelastung, sollten die gegenseitigen Einflüsse zumindest durch Schätzung soweit wie möglich herausgerechnet werden, um Redundanzen zu vermeiden.
- Konstante Merkmale, die für alle untersuchten Objekte dieselbe Ausprägung aufweisen, können als nicht trennungswirksam von vornherein außer acht gelassen werden.

Durch diese Hinweise kann das Problem der Merkmalsauswahl zwar nicht gelöst, die Wahl aber wenigstens etwas eingengt werden.

Ein weiteres Problem wird aufgeworfen, wenn in dem Merkmalsatz unterschiedliche Skalenniveaus² vertreten sind.

1 vgl. SNEATH, P.H.A., SOKAL, R.R.: Numerical Taxonomy. a.a.O., S. 103 ff.

2 vgl. BLEYMÜLLER, J., GEHLERT, G., GÜLICHER, H.: Statistik für Wirtschaftswissenschaftler. München 1979, S. 3 f. und die Kapitel 3 und 4, S. 13-23.

Die meisten iterativen Verfahren unterstellen, wie man den Zielfunktionen leicht ansehen kann¹, durchgängig metrisches Skalenniveau, wodurch sich die Verarbeitung ordinal- oder nominalskaliertter Merkmale mit diesen Verfahren eigentlich verbietet. Als Ausweg in solchen Fällen kann der Umweg über die Berechnung von Distanz- oder Ähnlichkeitsmaßen gelten, von denen eine Vielzahl für nominalskalierte, ordinalskalierte und sogar gemischte Datensätze existieren². Darauf kann dann entweder ein hierarchisches Verfahren oder ein Austauschverfahren für Distanzmatrizen³ angewendet werden.

Eine weitere interessante Möglichkeit bieten in diesem Zusammenhang die Verfahren der Mehrdimensionalen Skalierung⁴ (MDS): Sie ermöglichen es, aus einer Distanzmatrix **D**, die metrisches oder auch nur ordinales Skalenniveau aufweisen kann, eine Anordnung der Objekte in einem (in aller Regel) euklidischen Raum zu finden, sodaß die in diesem Raum gemessenen Distanzen, denen man nun - mit geringen Vorbehalten - metrisches Skalenniveau zubilligen kann, möglichst dieselbe Rangfolge wie die ursprünglichen Daten einhalten. Mit den Koordinaten bezüglich der Achsen dieses Raumes hat man gleichzeitig eine neue Datenmatrix **X** gefunden, der ebenfalls (mit geringen Vorbehalten) metrisches Skalenniveau zugemessen werden darf. Dieser Umweg würde dann wieder die Benutzung der Clusteranalyseverfahren für metrische Daten erlauben.

1 In den Zielfunktionen (2.34), (2.36) und (2.61) tauchen unmittelbar Differenzen und arithmetische Mittel auf, die erst bei metrischem Skalenniveau eine Information beinhalten; ebenso in den Streuungsmatrizen **T** und **W** bzw. **W_j** (s. (2.37) und (2.38)) von denen die Kriterien (2.48) und (2.64) und auch die Eigenwerte in (2.57) und (2.58) abhängen.

2 s. Abschnitt 3.2, S.63

3 s. z.B. STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 135 ff.

4 vgl. etwa KRUSKAL, J.: The Relationship between Multidimensional Scaling and Clustering. In: van RYZIN, J. (Hrsg.): Classification and Clustering. New York, San Francisco, London 1977, S. 17-44.

Sehr viel einfacher, formal gesehen unkorrekt, aber trotzdem nicht ohne weiteres von der Hand zu weisen ist demgegenüber das Verfahren, ordinalskalierte Daten einfach in Rangzahlen zu transformieren und diese wie metrische Merkmale zu behandeln. Zum einen sind empirischen Untersuchungen¹ zufolge in realistischen Situationen metrische Daten und zugehörige Ränge ohnehin meist hoch korreliert, sodaß der eventuelle Fehler nicht allzu groß wird, zum anderen werden auch bei der Berechnung von Distanzen und Ähnlichkeiten bei Ordinalskalen meist Rangziffern wie metrische Daten behandelt². Deshalb sind auf beiden Wegen die gleichen Verzerrungen zu erwarten. Für Nominalskalen allerdings dürfte die Verwendung spezieller Distanzmaße unumgänglich sein.

Die Frage, ob und wie stochastische Abhängigkeiten bzw. Korrelationen zwischen den gemessenen Variablen bereinigt werden sollen, wird unterschiedlich beantwortet. Berücksichtigt man nämlich mehrere hoch korrelierte Variable, hinter denen u.U. nur ein einziger relevanter Faktor³ steckt, so bekommt dieser Faktor dadurch, daß er mehrfach eingeht, ein stärkeres Gewicht als ein Faktor, der sich nur in einer einzigen Variablen auswirkt. Wird z.B. beim Clustern von Kfz der Faktor "Leistung" durch die beiden sehr hoch korrelierten Merkmale "Höchstgeschwindigkeit in km/h" und "Motorleistung in PS" repräsentiert (vgl. Abb. 3.2), so bekommt dieser Faktor bei der Clusterung nach Merkmalswerten (linker Teil von Abb. 3.2) erheblich mehr Gewicht als bei der Clusterung anhand der gleichgewichtigen Faktorenwerte. Daher unterscheiden sich die Cluster des Beispiels in Abb. 3.2 bei der Clusterung mit Merkmalen nur durch die Leistungskomponenten PS-Zahl und Höchstgeschwindigkeit, wogegen beim Clustern mit Faktoren (rechter Teil von Abb. 3.2) der Faktor Preis ebenfalls zur Unterscheidung beiträgt.

1 LABOVITZ, S.: The Assignment of Numbers to Rank Order Categories. American Sociological Review 35 (1970), S. 515 - 524

2 Vgl. z.B. die Formel (3.13) für den SPEARMAN'schen Rangkorrelationskoeffizienten auf S.65

3 zur Faktorenanalyse vgl. auch S. 8

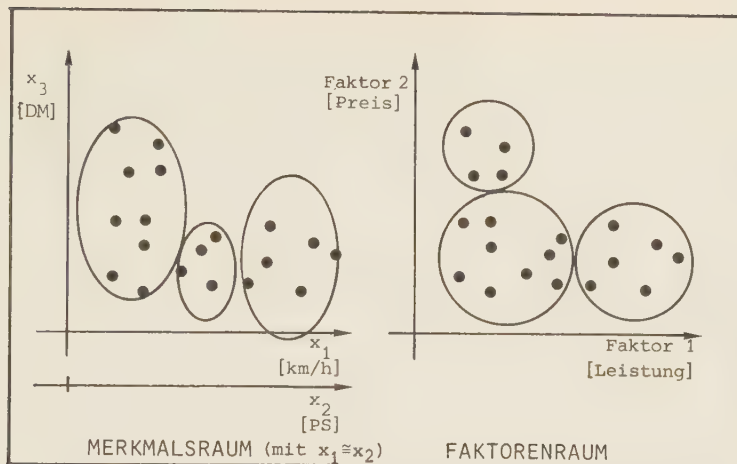


Abb. 3.2: Unterschiedliche Clusterung von KFZ im Merkmalsraum und im Faktorenraum

Der Einfluß des unterschiedlichen Gewichts von Faktoren, die aufgrund der zwangsläufig stark subjektiven Variablenauswahl von unterschiedlich vielen Variablen repräsentiert werden, kann weitgehend vermieden werden, wenn vor der Clusteranalyse eine Faktorenanalyse durchgeführt wird und für die Clusteranalyse nicht mehr die Variablenwerte, sondern die Faktorenwerte herangezogen werden. Dadurch bekommt jeder Faktor dasselbe Gewicht unabhängig davon, wieviele der ursprünglichen Variablen von ihm abhängen. Die so erzielte Gleichgewichtigkeit der Faktoren ist aber andererseits auch wieder angreifbar, weil man einen Faktor, von dem mehrere der ja bereits als relevant akzeptierten Variablen abhängen, tatsächlich als wichtiger ansehen kann als einen Faktor, der nur von einer einzigen Variablen repräsentiert wird. Von daher wäre es sinnvoll, die ursprünglichen Daten ohne vorherige Faktorenanalyse zu verwenden, weil die einflußreichen Faktoren dann auch entsprechend mehr Gewicht in der Clusteranalyse erhalten.

Das Problem der angemessenen Gewichtung verschiedener Merkmale wird leider nicht nur von inhaltlichen Überlegungen, sondern auch von reinen Äußerlichkeiten berührt. Die angestrebte Proportionalität zwischen der sachlichen Bedeutung eines Merkmals und seinem Einfluß auf das Klassifikationsergebnis kann nämlich bereits von der Feinheit der Merkmalsskala erheblich gestört werden. Ein Merkmal mit zahlenmäßig sehr großen Ausprägungen, z.B. die Einwohner von Städten, hat mehr Einfluß als ein Merkmal mit zahlenmäßig kleinen Ausprägungen wie die Arbeitslosenquote in %, weil auch die Differenzen beim ersten Merkmal zwangsläufig wesentlich größer als beim zweiten sind. Meist wird aber gerade an den Differenzen die Ähnlichkeit bzw. Unähnlichkeit der Objekte gemessen. Daher kann ein einzelnes Merkmal dominierend für das gesamte Klassifikationsergebnis wirken, wenn es lediglich in einer besonders feinen Maßeinheit gemessen wird. So verliert z.B. das Merkmal "Preis" gegenüber dem Merkmal "Höchstgeschwindigkeit" im Beispiel in Abb. 3.3 an Einfluß, wenn es statt in DM (linker Teil der Abb. 3.3) in \$ gemessen wird (rechter Teil der Abb. 3.3), weil dadurch die Preisdifferenzen zahlenmäßig kleiner werden. Die Cluster unterscheiden sich dann nur noch durch die unterschiedlichen Höchstgeschwindigkeiten.

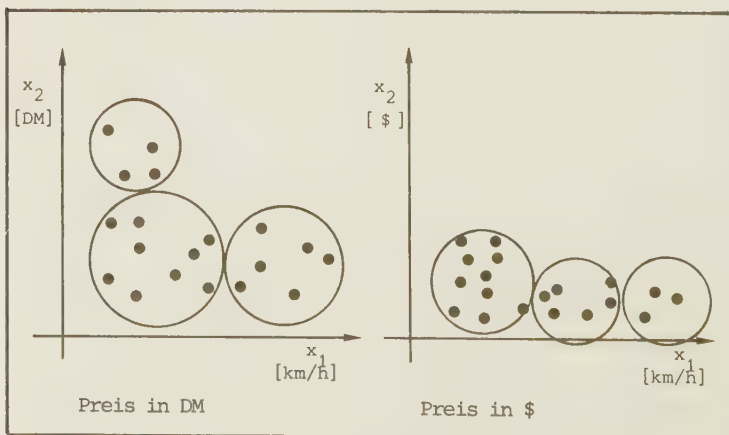


Abb. 3.3: Unterschiedliche Clusterung von KFZ bei Messung eines Merkmals in verschieden feinen Maßeinheiten

Die Vergleichbarkeit der Meßwerte in dem Sinne, daß alle Merkmale dasselbe Gewicht bekommen, kann durch eine Reihe von normierenden Transformationen hergestellt werden, die meist ein Streumaß wie die Varianz oder die Spannweite vereinheitlichen. Zu den am häufigsten gebrauchten Transformationen dieser Art gehören die folgenden¹:

- (1) Bei der Standardisierung oder z-Transformation werden die Mittelwerte der Variablen auf 0 und die Standardabweichungen auf 1 normiert durch die Abbildung

$$x_{ij} \mapsto \frac{x_{ij} - \bar{x}_j}{s_j} \quad , \quad (3.1)$$

wobei $\bar{x}_j$ der Mittelwert der j-ten Variablen, also die j-te Komponente des Mittelwertvektors $\bar{\mathbf{x}}$ und s_j die durch

$$s_j = \left[\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right]^{1/2} \quad (3.2)$$

definierte Standardabweichung² der j-ten Variablen ist. Im Rahmen dieser Arbeit wurde für Standardisierungen das Unterprogramm ZTRANS aus der Programmsammlung von STEINHAUSEN/LANGER benutzt³.

-
- 1 Vgl. VOGEL, F.: Subjektivitäten bei der Klassifikation von Einheiten. In: Proceedings in Operations Research. Jahrestagung DGOR 1977, Würzburg, Wien 1978, S. 105 - 129
- 2 In statistischen Lehrbüchern wird die Varianz einer Stichprobe teilweise auch als $s_j^2 = \frac{1}{n} \sum (x_{ij} - \bar{x}_j)^2$ definiert, woraus sich eine erwartungstreue Schätzfunktion für die Varianz σ_j^2 der j-ten Variablen (bei stochastischer Unabhängigkeit und identischer Verteilung aller Einzelbeobachtungen) als $\sigma_j^2 = \frac{n}{n-1} s_j^2$ ergibt. Häufig wird sie jedoch direkt als erwartungstreue Schätzfunktion entsprechend (3.2) definiert. Für die Standardisierung bei Clusteranalysen ist der Unterschied jedoch ohne Bedeutung, weil er alle Variablen gleichermaßen betrifft und daher keine Gewichtsverschiebung verursacht.
- 3 STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 72

Bei der Verwendung der L_p -Metriken (s.(2.60)) mit $p \in \mathbb{N}$ kann es zweckmäßig sein, statt einer Normierung der Varianz, die auch als 2. zentrales Moment¹ bezeichnet wird, gerade die Momente der Ordnung p zu vereinheitlichen; dazu braucht in (3.1) lediglich s_j durch

$$s_j^{(p)} = \left[\frac{1}{n} \sum_{i=1}^n |x_{ij} - \bar{x}_j|^p \right]^{1/p} \quad (3.3)$$

ersetzt zu werden.

(2) Mit der [0,1]-Normierung

$$x_{ij} \mapsto \frac{x_{ij} - \min_{1 \leq i \leq n} \{x_{ij}\}}{\max_{1 \leq i \leq n} \{x_{ij}\} - \min_{1 \leq i \leq n} \{x_{ij}\}} \quad (3.4)$$

werden die Beobachtungswerte so transformiert, daß sie im Intervall $[0,1]$ liegen und die Spannweite 1 aufweisen.

(3) Mit der Median-Normierung wird schließlich eine Verschiebung des Medians auf 0 und eine Mittlere Absolute Abweichung vom Median von 1 erzielt; sie lautet

$$x_{ij} \mapsto \frac{x_{ij} - \tilde{x}_j}{\frac{1}{n} \sum_{i=1}^n |x_{ij} - \tilde{x}_j|} \quad , \quad (3.5)$$

wobei $\tilde{x}_j$ der Median der j -ten Variablen ist².

1 vgl. KREYSZIG, E.: Statistische Methoden und ihre Anwendungen. Göttingen, 6. Aufl. 1977, S. 92 f.

2 Der Median der Beobachtungswerte $x_{1j}, \dots, x_{nj}$ der j -ten Variablen wird aus der Reihe der der Größe nach geordneten Werte $x_{[1]j}, \dots, x_{[n]j}$ bestimmt als

$$x_j = \begin{cases} x_{\left[\frac{n+1}{2}\right]j} & n \text{ ungerade} \\ \frac{1}{2} \left(x_{\left[\frac{n}{2}\right]j} + x_{\left[\frac{n}{2}+1\right]j} \right) & n \text{ gerade} \end{cases} \quad \text{für}$$

vgl. BLEYMÜLLER, J., GEHLERT, G.: Statistische Formeln und Tabellen. München 1979, S. 11.

Normalerweise ist eine Vorhersage darüber, in welcher Weise und wie stark sich eine spezielle Form der Normierung bei einem bestimmten Datensatz auf die Endergebnisse auswirken wird, kaum möglich. In einer Studie von VOGEL konnte lediglich festgestellt werden, daß die Klassifikationsergebnisse bei einer Gruppe von Normierungen, zu denen die o.g. Standardisierung (3.1) und die $[0,1]$ -Normierung (3.4) gehören, recht gut übereinstimmen¹. Die in praktischen Anwendungen zu verzeichnende starke Bevorzugung der Standardisierung scheint also durch diese Untersuchung gerechtfertigt zu werden, da sie gleichzeitig als Stellvertreter für eine ganze Gruppe von Transformationen mit ähnlicher Wirkung gelten kann.

Derartige Normierungen können wie gesagt die rein zufälligen, aus der unterschiedlichen Feinheit der verschiedenen Merkmalsskalen resultierenden Ungleichgewichte zwischen den Variablen beseitigen. Die Frage, welche Gewichte stattdessen besser geeignet sind, um damit brauchbare Ergebnisse zu produzieren, wird dadurch jedoch nicht beantwortet, denn es gibt keinen Grund anzunehmen, daß die Gleichgewichtigkeit der Variablen sich vor anderen möglichen Gewichtsverteilungen auszeichnet. Es wäre im Gegenteil sehr ungewöhnlich, wenn alle Variablen für das Untersuchungsziel genau die gleiche Bedeutung hätten. Vor einer Weiterverarbeitung der Daten müßte deshalb, nachdem durch eine Normierung die maßstabsbedingte zufällige Gewichtung beseitigt worden ist, eine neue Gewichtung

$$x_{ij} \longrightarrow g_j x_{ij} \quad (3.6)$$

vorgenommen werden mit Faktoren $g_j \in \mathbb{R}^+$, $1 \leq j \leq m$, die der

1 s. VOGEL, F.: Subjektivitäten bei der Klassifikation von Einheiten. a.a.O. In dieser Studie werden die Auswirkungen mehrerer Transformationen auf einen Beispieldatensatz beobachtet, der jeweils mit dem Verfahren von WARD (s.S. 30) hierarchisch geclustert wird. Da die beobachteten Unterschiede aber sowohl vom Datensatz als auch vom Clusteranalysealgorithmus abhängen können, ist nicht klar, inwiefern die Beobachtungen verallgemeinert werden können.

Bedeutung der jeweiligen Merkmale entsprechen.

Leider ist aber i.d.R. nicht im voraus zu erkennen, welche Variablen am meisten zur Auftrennung der Cluster beitragen. Dies kann man erst anhand der fertigen Partition herausfinden, sodaß im Nachhinein wohl die entscheidenden Variablen erkannt und bei einer eventuellen Wiederholung der Klassifizierung stärker gewichtet werden können, um die in der vorigen Lösung unter Umständen nur undeutlich ausgeprägte Gruppenstruktur besser herauszuarbeiten. Aber selbst in diesem iterativen Prozeß bestehen noch Möglichkeiten, nach rein subjektivem Ermessen bestimmte Merkmale besonders zu bevorzugen, indem z.B. die von ihnen bewirkten Aufteilungen bzw. Zusammenfassungen als charakteristisch und andere als artefiziell angesehen werden. Für die Verteilung der Gewichte vor dem ersten Versuch der Klassifizierung gibt es überhaupt keinen methodischen Anhaltspunkt; die Gewichtung der Merkmale muß daher genauso dem Ermessen des Anwenders überlassen bleiben wie schon zuvor die Auswahl¹. Als Ausweg aus diesem Dilemma werden häufig die skaleninvarianten Methoden² angesehen, die sich von einer Änderung der Maßeinheiten, also auch von einer Gewichtung nicht beeinflussen lassen; tatsächlich nehmen aber solche Verfahren indirekt nur eine Gleichgewichtung der Variablen vor, sodaß man die Frage nach den richtigen Gewichten dadurch nicht als beantwortet, sondern lediglich als umgangen ansehen kann.

1 Die Auswahl kann ohnehin als Grenzfall der Gewichtung angesehen werden, denn durch ein Gewicht von $g_j = 0$ wird das Merkmal j völlig eliminiert.

2 Vgl. Abschnitt 3.2 über die Auswahl der Verfahren.

3.2 DISTANZ- UND ÄHNLICHKEITSMABE

Für die meisten hierarchischen Clusteranalyseverfahren müssen die Daten in Form einer Distanzmatrix **D** bereitgestellt werden.

- Diese kann in einigen Fällen direkt erhoben werden¹, sodaß eine weitere Aufbereitung nicht notwendig ist.
- Falls als Ausgangsdaten eine Datenmatrix **X** zur Verfügung steht, können daraus Distanzen d_{ik} oder Ähnlichkeiten s_{ik} berechnet werden.
- Gegebenenfalls kann eine Ähnlichkeitsmatrix durch geeignete Transformationen noch in eine Distanzmatrix umgewandelt werden.

Zur Umwandlung einer Ähnlichkeitsmatrix in eine Distanzmatrix eignen sich solche Transformationen, die gewährleisten, daß aus den Eigenschaften von Ähnlichkeitsmaßen die entsprechenden Eigenschaften von Distanzmaßen² hervorgehen. Transformationen dieser Art sind z.B.:

$$d_{ik} = s_{\max} - s_{ik} \quad , \quad (3.7)$$

$$d_{ik} = \log(s_{\max}) - \log(s_{ik}) \quad , \quad (3.8)$$

$$d_{ik} = \sqrt{2(1-s_{ik})} \quad \text{für } -1 \leq s_{ik} \leq 1 \quad . \quad (3.9)$$

-
- 1 Häufig sind die untersuchten Objekte nicht gegenständlich, sondern z.B. Meinungsbilder (etwa von Konsumenten über Produkte). Dann ist es oft sogar unmöglich, die darin enthaltenen Variablen zu bestimmen und zu messen. Auf jeden Fall ist es in solchen Situationen einfacher, sofort nach der Ähnlichkeit oder Unähnlichkeit der Objekte zu fragen; s. z.B. NABER, A.: Multivariate Analyse von Produktmärkten. Dissertation, Münster 1979, S. 19 ff.
 - 2 Zu den definierenden Eigenschaften von Distanz- und Ähnlichkeitsmaßen vgl. S. 20.

Bei der Bestimmung einer Distanz- oder Ähnlichkeitsmatrix aus einer Datenmatrix ergibt sich noch einmal eine Reihe von Möglichkeiten, durch die Wahl entsprechender Maße auf die bereits bei der Aufbereitung von **X** diskutierten Fragen wie Merkmalsgewichtung, Berücksichtigung von Korrelationen usw. einzugehen¹. Der dominierende Gesichtspunkt für die Wahl eines Maßes ist jedoch das Skalenniveau der Merkmale.

Für einen Merkmalssatz, der ausschließlich metrisch skalierte Merkmale enthält, gibt es eine außerordentlich große Zahl von Distanz- und Ähnlichkeitsmaßen, von denen die gebräuchlichsten zu den folgenden Typen gehören:

- Die L_p -Metriken oder MINKOWSKI-Metriken messen den Abstand zweier Objekte i und k als

$$d_{ik}^p = \left[\sum_{j=1}^m |x_{ij} - x_{kj}|^p \right]^{1/p}, \quad (3.10)$$

wobei p eine positive reelle Konstante ist.

Als Spezialfälle ergeben sich daraus mit $p = 1$ die sog. City-Block-Distanz und

mit $p = 2$ die bei weitem bekannteste Euklidische Distanz.

Als Grenzfall, nämlich für $p \rightarrow \infty$, erhält man die sog. Maximumsmetrik

$$d_{ik}^\infty = \max_{1 \leq j \leq m} |x_{ij} - x_{kj}|, \quad (3.11)$$

bei der lediglich die größte Merkmalsdifferenz zur Abstandsbestimmung herangezogen wird.

- Durch Potenzieren der Merkmalsdifferenzen kann z.B. die Randlage eines Ausreißers besonders deutlich hervorgehoben werden. Das am meisten benutzte Distanzmaß dieser

¹ Eine sehr ausführliche Zusammenstellung bekannter Maße und der Möglichkeiten, Maße zu konstruieren, findet sich z.B. bei BOCK, H.H.: Automatische Klassifikation. a.a.O., S. 24-79. Tabellarische Übersichten der verbreitetsten Maße nach den Skalenniveaus der Merkmalsätze getrennt enthält die Arbeit von CLEMENS, U.: Mehrdimensionale Skalierung bei gemischten Daten. Dipl.-Arbeit, Münster 1980.

Art ist die Quadrierte Euklidische Distanz

$$d_{ik} = \sum_{j=1}^m (x_{ij} - x_{kj})^2 \quad . \quad (3.12)$$

Es ist natürlich ohne weiteres möglich, in die bisher genannten Distanzmaße für jede Variable einen Gewichtungsfaktor $g_j \in \mathbb{R}^+$, $1 \leq j \leq m$, einzufügen, wenn die Merkmale nicht gleichgewichtig in die Distanzberechnung eingehen sollen.

- Nicht nur eine Gewichtung der Variablen, sondern auch eine Berücksichtigung der Korrelationen durch eine geeignete Verzerrung des Merkmalsraumes ermöglichen die durch Quadratische Formen¹ definierten Distanzmaße des Typs

$$d_{ik} = \left[(\mathbf{x}_i - \mathbf{x}_k) \mathbf{A} (\mathbf{x}_i - \mathbf{x}_k)^t \right]^{1/2} \quad (3.13)$$

mit einer positiv definiten symmetrischen Matrix $\mathbf{A} \in \mathbb{R}^{(m,m)}$, die Art und Ausmaß der Verzerrung des Merkmalsraumes bestimmt. Beispielsweise bewirkt die Matrix $\mathbf{A} = \begin{pmatrix} 1.00 & 0.81 \\ 0.81 & 1.00 \end{pmatrix}$, daß alle in Abb. 3.4 eingezeichneten Punkte bei Abstandsmessung nach (3.13) den gleichen Abstand von 1.0 zu dem Punkt $\mathbf{x}_0$ haben, obwohl ihre euklidischen Abstände zwischen 0.75 und 2.30 schwanken.

Mit $\mathbf{A} = \mathbf{I}_m$ erhält man als Spezialfall auch hier wieder die Euklidische Distanz.

Ist $\mathbf{A}$ eine Diagonalmatrix, die auf der Diagonalen Faktoren $g_j \in \mathbb{R}^+$, $1 \leq j \leq m$, für jede Variable enthält, so entsteht daraus eine Gewichtete Euklidische Distanz.

Besonders interessant ist die Verwendung der Inversen der Streuungsmatrix, also $\mathbf{T}^{-1}$ als Matrix der quadratischen Form. Das so entstehende, als MAHALANOBIS-Distanz

1 Eine quadratische Form des $\mathbb{R}^m$ ist eine durch $\mathbf{x} \mapsto \mathbf{x} \mathbf{A} \mathbf{x}^t$ mit $\mathbf{x} \in \mathbb{R}^m$ und eine symmetrische Matrix $\mathbf{A} \in \mathbb{R}^{(m,m)}$ definierte Abbildung des $\mathbb{R}^m$ in den $\mathbb{R}^1$. Vgl. z.B. ZURMÜHL, R.: Matrizen und ihre technischen Anwendungen. Berlin, Göttingen, Heidelberg, 4. neu bearb. Aufl. 1964, S. 125 ff.

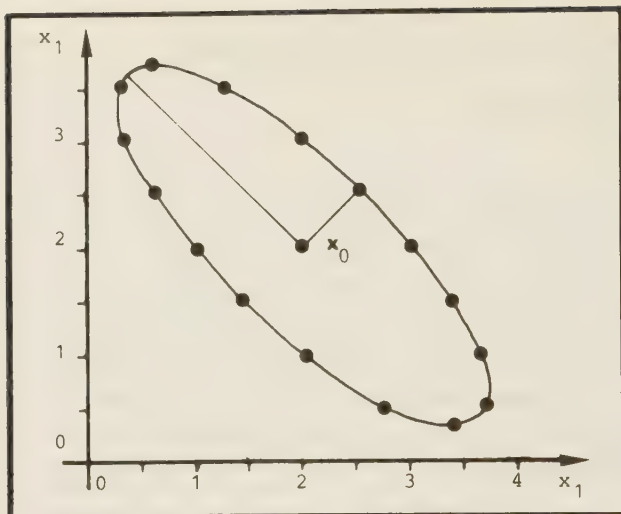


Abb. 3.4: Punkte gleichen Abstands von x_0 bei Distanzmessung nach (3.13)

bekannte Maß eliminiert sowohl Ungleichgewichte als auch Korrelationen der Variablen¹ und paßt sich der Gestalt der Punktwolke im Merkmalsraum insofern an, als zur Berechnung von d_{ik} nicht nur die Merkmalsausprägungen der Objekte i und k , sondern auch die Lage aller anderen Objekte zueinander herangezogen wird.

- Häufig wird auch das Konzept, stochastische Abhängigkeit von Zufallsvariablen durch Korrelationskoeffizienten zu messen, auf die Ähnlichkeitsmessung von Objekten übertragen. Auf rein metrischem Skalenniveau kann z.B. der Produktmomentkorrelationskoeffizient zur Definition des Ähnlichkeitsmaßes

$$s_{ik} = \frac{\sum_{j=1}^m (x_{ij} - \bar{x}_i) (x_{kj} - \bar{x}_k)}{\sqrt{\sum_{j=1}^m (x_{ij} - \bar{x}_i)^2 \sum_{j=1}^m (x_{kj} - \bar{x}_k)^2}}^{1/2} \quad (3.14)$$

herangezogen werden, wobei

¹ zum Beweis s. z.B. BOCK, H.H.: Automatische Klassifikation. a.a.O., S. 40 ff.

$$\bar{x}_{i.} = \frac{1}{m} \sum_{j=1}^m x_{ij} \quad (3.15)$$

der Mittelwert der Beobachtungswerte des i-ten Objektes ist. Schon wegen der unterschiedlichen Dimensionen der einzelnen Meßwerte kann $\bar{x}_{i.}$ und damit auch das Maß (3.14) wohl nur dann sinnvoll sein, wenn die Meßwerte $x_{i1}, \dots, x_{im}$ durch normierende Transformationen schon vorher vergleichbar gemacht worden sind.

Während für metrisch skalierte Merkmalssätze eine recht umfangreiche Palette von Distanz- und Ähnlichkeitsmaßen entwickelt worden ist, gibt es nur recht wenige spezielle Maße für ordinalskalierte Variable.

Dazu gehören hauptsächlich die in der parameterfreien Statistik entwickelten Prüfgrößen, die allein mit Rangziffern auskommen. Nimmt man an, daß die x_{ij} , $1 \leq j \leq m$ bereits Rangziffern sind, daß also x_{ij} der Rang des i-ten Objektes in der Reihenfolge der Ausprägungen des j-ten Merkmals ist, so kann z.B. ein Ähnlichkeitsmaß durch den SPEARMAN'schen Rangkorrelationskoeffizienten¹ wie folgt definiert werden:

$$s_{ik} = 1 - \frac{6 \sum_{j=1}^m (x_{ij} - x_{kj})^2}{m^3 - m} \quad (3.16)$$

Diese Größe dient bei der Untersuchung von Variablen zum Test auf stochastische Unabhängigkeit.

1 s. z.B. LEHMANN, E.L.: Nonparametrics, Statistical Methods based on Ranks. San Francisco 1975, S. 297 ff. Hier werden die Objekte wie Merkmale behandelt und hohe Korrelation wie beim Produktmomentkorrelationskoeffizienten (3.14) als Ähnlichkeit interpretiert.

Häufiger als die Benutzung eines speziellen Maßes für Ordinalskalen, die ohnehin nur selten allein vorkommen, ist jedoch das Ausweichen auf höheres Skalenniveau, indem die Rangziffern einfach wie metrische Merkmalsausprägungen behandelt werden.

Objekte mit nominalskalierten Merkmalen lassen sich nur noch dadurch vergleichen, daß überprüft wird, ob die Ausprägungen der Merkmale übereinstimmen oder nicht. Im einfachsten Fall, das sind Merkmale mit lediglich zwei Ausprägungen 1 und 0 (z.B. Besitz und Nichtbesitz eines Kennzeichens), hängen daher alle skalengerechten Distanz- und Ähnlichkeitsmaße von den vier Zahlen $a, b, c, e \in \mathbb{N}_{ab}$, die sich bei der Gegenüberstellung zweier Objekte etwa aus der folgenden Vierfeldertafel¹ ergeben:

Merkmalsausprägungen am Objekt k	Merkmalsausprägungen am Objekt i		Σ
	1	0	
1	a	b	a + b
0	c	e	c + e
Σ	a + c	b + e	m

Tab. 3.1: Anzahl der übereinstimmenden und nichtübereinstimmenden Merkmalsausprägungen bei zwei Objekten mit dichotomen Merkmalen.

1 S. auch STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der numerischen Klassifikation. a.a.O., S. 53.

In Tabelle 3.1 sind

- a und e die Zahlen derjenigen Merkmale, deren Ausprägungen bei beiden Objekten übereinstimmen, wobei oft a die Zahl der bei beiden vorhandenen Kennzeichen und e die Zahl der bei beiden nicht vorhandenen Kennzeichen ist,
- b und c die Zahlen derjenigen Merkmale, in denen die Objekte nicht übereinstimmen.

Aus diesen Zahlen von Übereinstimmungen und Nichtübereinstimmungen läßt sich nun eine ganze Reihe von Ähnlichkeitsmaßen konstruieren, von denen die meistgebrauchten in eine der beiden folgenden Gruppen fallen¹:

- Durch

$$s_{ik} = \frac{a+ae}{a+ae+\beta(b+c)} \quad (3.17)$$

mit $0 \leq \alpha \leq 1$ und $\beta > 0$ kann eine ganze Schar von Ähnlichkeitsmaßen definiert werden, wobei durch die Wahl von α das Gewicht der gemeinsam nicht vorhandenen Kennzeichen e verringert werden kann und durch die Wahl von β der Einfluß der oft zwangsläufig sehr zahlreichen Nichtübereinstimmungen zurückgedrängt werden kann.

Unter anderem erhält man mit $\alpha = 0$ und $\beta = 1$ den Koeffizienten von TANIMOTO (bzw. JACCARD, auch Similarity-S), der nur die gemeinsam vorhandenen, aber nicht die gemeinsam nicht vorhandenen Kennzeichen berücksichtigt.

1 Eine größere Zahl von Maßen findet man z.B. bei CLEMENS, U.: Mehrdimensionale Skalierung bei gemischten Daten. a.a.O., in tabellarischen Übersichten auf S. 34 u. 38.

Mit $\alpha = 1$ und $\beta = 1$ erhält man den Simple Matching Coefficient, der beide Ausprägungen dichotomer Merkmale als gleichwertig ansieht.

- Weiterhin werden wieder einige Korrelationskoeffizienten und deren Modifikationen zur Ähnlichkeitsmessung verwendet wie z.B. der durch

$$s_{ik} = \frac{ae - bc}{(a+e)(b+e)(a+b)(c+e)} \quad (3.18)$$

definierte Φ -Koeffizient, der ein Analogon zum Produktmomentkorrelationskoeffizienten (3.14) und zum SPEARMAN'schen Rangkorrelationskoeffizienten (3.16) ist und auch als Prüfgröße eines χ^2 -Tests auf "Unabhängigkeit"¹ dient.

Für nominale Merkmale mit mehr als zwei Ausprägungen bietet sich immer die Möglichkeit, sie in mehrere dichotome Merkmale zu zerlegen. Der einfachste Weg wäre, für jede Ausprägung je eine Hilfsvariable zu definieren, die den Wert 1 bei Vorliegen dieser Ausprägung und sonst den Wert 0 annimmt. Für diese Hilfsvariablen können dann die bereits bekannten Maße für dichotome Merkmale verwendet werden². Die gleichen Ergebnisse wie bei der Zerlegung und anschließender Verwendung eines aus (3.17) abgeleiteten Maßes erhält man aber einfacher, wenn stattdessen das jeweilige Ähnlichkeitsmaß verallgemeinert wird. Dazu muß lediglich in der Formel (3.17)

1 Auch hier werden wieder wie in (3.14) und (3.16) Objekte wie Variable behandelt und hohe Korrelation wird als Ähnlichkeit interpretiert.

2 Dieses Vorgehen ist nicht unkritisch, weil dadurch eine mehrwertige nominalskalierte Variable das mehrfache Gewicht einer Variablen erhält, die von vornherein zweiwertig ist. Vgl. z.B. SODEUR, W.: Empirische Verfahren zur Klassifikation. Stuttgart, 1974, S. 62 f. u. 66 ff.

- a gegen m_a , die Anzahl aller Merkmalsausprägungen, die bei beiden Objekten vorliegen,
- e gegen m_e , die Anzahl aller Merkmalsausprägungen, die bei beiden Objekten nicht vorliegen, und
- b+c gegen m_{b+c} , die Anzahl derjenigen Merkmalsausprägungen, die bei einem Objekt vorliegen und beim anderen nicht,

ausgetauscht werden.

- Der so verallgemeinerte Simple Matching Coefficient beispielsweise würde dann als

$$s_{ik} = \frac{m_a + m_e}{m_a + m_e + m_{b+c}} = 1 - \frac{2(m-m_a)}{p} \quad (3.19)$$

berechnet, wobei p die Gesamtzahl der Ausprägungen aller Merkmale ist; dieses Maß ist äquivalent¹ zu dem anschaulich naheliegenden Maß

$$s_{ik} = \frac{m_a}{m} \quad , \quad (3.20)$$

das einfach den Anteil der übereinstimmenden Merkmale zur Bestimmung der Ähnlichkeit heranzieht.

Häufig enthalten Merkmalssätze nicht ausschließlich Merkmale mit demselben Skalenniveau, sondern eine Mischung aus nominal-, ordinal- und metrisch skalierten Merkmalen, sodaß keine der bisher angeführten Möglichkeiten, Distanzen oder Ähnlichkeiten zu berechnen, geeignet ist, sämtliche Merkmale zugleich zu verarbeiten. Für diesen Fall sind einige Konzepte entwickelt worden, deren einsichtigstes darin be-

1 (3.20) geht aus (3.19) lediglich über die Lineare Transformation

$s_{ik} \mapsto (2 - \frac{p}{m}) - \frac{p}{m} s_{ik}$ hervor.

steht, ein gemeinsames Distanzmaß aus je einem Distanzmaß für die nominalen, ordinalen und metrischen Merkmale zusammenzusetzen¹. Dazu werden die Beobachtungsvektoren $\mathbf{x}_i$, $1 \leq i \leq n$, jeweils in einen nominalskalierten Teil $\mathbf{x}_i^{(n)}$, einen ordinalskalierten Teil $\mathbf{x}_i^{(o)}$ und einen metrisch skalierten Teil $\mathbf{x}_i^{(m)}$ zerlegt, sodaß

$$\mathbf{x}_i = (\mathbf{x}_i^{(n)}, \mathbf{x}_i^{(o)}, \mathbf{x}_i^{(m)}) \quad (3.21)$$

gilt und die Distanz $d_{ik} = d(\mathbf{x}_i, \mathbf{x}_k)$ als

$$\begin{aligned} d(\mathbf{x}_i, \mathbf{x}_k) &= g^{(n)} d^{(n)}(\mathbf{x}_i^{(n)}, \mathbf{x}_k^{(n)}) \\ &+ g^{(o)} d^{(o)}(\mathbf{x}_i^{(o)}, \mathbf{x}_k^{(o)}) \\ &+ g^{(m)} d^{(m)}(\mathbf{x}_i^{(m)}, \mathbf{x}_k^{(m)}) \end{aligned} \quad (3.22)$$

bestimmt werden kann, wobei $d^{(n)}$, $d^{(o)}$ und $d^{(m)}$ Distanzmaße für nominal-, ordinal- bzw. metrisch skalierte Merkmale sind. Mit den Gewichten $g^{(n)}$, $g^{(o)}$ und $g^{(m)} \in \mathbb{R}^+$ können die durch die Verschiedenartigkeit der Maße bedingten Ungleichgewichte aufgefangen werden.

Der einfachste Weg der Verarbeitung gemischter Merkmalsätze ist allerdings, alle nominalen Variablen in dichotome Hilfsvariablen zu zerlegen, ordinale Merkmalsausprägungen in Rangziffern umzuwandeln und die Gewichte aller Variablen durch eine Normierung² auszugleichen. Anschließend können alle Variablen gemeinsam mit einem Distanzmaß für metrische Skalen weiterverarbeitet werden³. Dieses Ver-

1 Genauso werden natürlich auch Ähnlichkeitsmaße zusammengesetzt. S. z.B. GOWER, J.C.: A General Coefficient of Similarity and some of its Properties. Biometrics 27 (1971), S. 857-874.

2 s.S. 59

3 z.B. wird bei PARKS, J.M.: Classification of Mixed Mode Data by R-Mode Factor Analysis and Q-Mode Cluster Analysis on Distance Functions. In: COLE, A.J. (Hrsg.): Numerical Taxonomy. London, New York 1969, S. 216-219, die Kombination der [0,1]-Normierung (3.4) mit der Euklidischen Distanz ((3.10) mit $p = 2$) vorgeschlagen.

fahren entspricht zwar einer Niveauproggression ohne die dazu notwendigen Zusatzinformationen und ist daher eigentlich unzulässig. In Anbetracht der Tatsache jedoch, daß auch bei den speziellen Distanz- und Ähnlichkeitsmaßen für nichtmetrische Skalen entweder die Merkmalsausprägungen wie metrisch behandelt werden oder die resultierenden Maße wieder nur ordinales Skalenniveau haben und daher mit Maßen für metrische Skalen nicht vergleichbar sind, kann diese scheinbar oberflächliche Vorgehensweise kaum als schlechter angesehen werden als z.B. die differenzierte Berücksichtigung der Skalenniveaus in (3.22).

3.3 DIE KOMBINATION VON VERFAHRENSBESTANDTEILEN

In den vorangegangenen Abschnitten wurden zu den mit der Durchführung einer Clusteranalyse verbundenen Einzelproblemen jeweils einige der wichtigsten Lösungsmöglichkeiten geschildert. Die Liste dieser Möglichkeiten erhebt keinerlei Anspruch auf Vollständigkeit, sondern dient eher zur Demonstration der Vielfalt bereits vorhandener Techniken und der Methoden, situationsspezifische Lösungen zu entwickeln. Trotz der Beschränkung auf wenige, wenn auch häufig benutzte Bausteine auf jeder Stufe des Klassifikationsvorganges dürfte aber bereits deutlich geworden sein, daß durch die Kombination der einzelnen Schritte von der Datenaufbereitung über die Berechnung der Distanzen bis zur Auswahl eines Algorithmus bzw. einer Zielfunktion eine äußerst unübersichtliche Fülle von unterschiedlichen Vorgehensweisen entsteht. In der Praxis wird daher weniger oft die Frage auftauchen, ob es für eine bestimmte Situation eine geeignete Bausteinkombination gibt, als die Notwendigkeit, aus der Reichhaltigkeit des Angebots ungeeignete Verfahren auszuschneiden.

Als Orientierungshilfe werden deshalb im folgenden die einzelnen Bausteine mitsamt ihren für die Auswahl wichtigen Eigenschaften tabellarisch aufgelistet, um auf Basis dieser Tabellen anschließend einen Wegweiser in Form eines Programmablaufplanes zu erstellen (s. Abb. 3.5). Dieser kann natürlich kaum als verbindlich angesehen werden, wenn man bedenkt, daß viele Techniken - wie z.B. die hierarchischen Clusteranalysealgorithmen - heuristisch motiviert sind und nicht auf bestimmten zwingenden Modellvoraussetzungen aufbauen; in einigen Situationen stehen auch offenbar ausschließlich unbefriedigende Wahlmöglichkeiten zur Verfügung, z.B. für die Distanzmessung bei gemischten Merkmalssätzen. Daher führt auch dieser Plan i.d.R. nicht

zu einer eindeutigen Bausteinkombination, sondern scheidet lediglich eine Anzahl unsinniger Kombinationen aus, wie beispielsweise die Folge aus Standardisierung, Gewichtung und anschließende Clusterung mit einer skaleninvarianten Zielfunktion, die die Gewichtung wieder zunichte macht.

Für die Auswahl von hierarchischen Algorithmen und Zielfunktionen für iterative Algorithmen sind hauptsächlich Invarianzeigenschaften ausschlaggebend, weil sie mit dem Skalenniveau der Eingabedaten, das zumeist als bekannt vorausgesetzt werden kann, korrespondieren (Tab. 3.2 und 3.3).

Die Zielfunktionen für iterative Verfahren sind alle für metrische Merkmale gedacht, ferner sind die von einem Austauschalgorithmus erzeugten (suboptimalen) Lösungen immer von der Startpartition und damit von der Eingabereihenfolge abhängig; diese Punkte sind deshalb nicht entscheidungsrelevant (Tab. 3.3).

Für Distanzmaße sind wiederum die Invarianzeigenschaften bzw. Skalenniveaus der Merkmale interessant (Tab. 3.5).

Name	Formel	Skalenniveau der Distanzmatrix	Invarianz der Ergebnisse bei ¹			linearen Transf.	Bemerkung
			monotonen Transf.	Translationen	Ähnlichkeitstr.		
Single Linkage	2.17	ordinal	ja	ja	ja	ja	neigt zu Kettenbildung
Complete Linkage	2.18	ordinal	ja	ja	ja	ja	-
Average Linkage	2.19	metrisch	nein	ja	ja	ja	-
Weighted Average Linkage	2.21	metrisch	nein	ja	ja	ja	-
Median-Verfahren	2.23	metrisch	nein	nein	ja	nein	für quadrierte euklidische Distanzen (2.22)
Centroid-Verfahren	2.26	metrisch	nein	nein	ja	nein	
WARDs Verfahren	2.30	metrisch	nein	ja	ja	ja	
Flexible Strategie	2.31	metrisch	nein	nein	ja	nein	minimiert das Varianzkr. (2.27) bei quadr. eukl. Distanzen (2.22) $\alpha_i = \alpha_j = \alpha, \gamma = 0$ und $\beta = 1 - 2\alpha$

Tab. 3.2: Eigenschaften hierarchischer Verfahren

1 Eine Transformation f der Elemente der Distanzmatrix heißt

monoton $\Leftrightarrow [d < d' \Rightarrow f(d) < f(d')] \quad \forall d, d'$,

Translation $\Leftrightarrow f(d) = d + b \quad \forall d$ mit $b \in \mathbb{R}$,

Ähnlichkeitstransformation $\Leftrightarrow f(d) = ad \quad \forall d$ mit $a \in \mathbb{R}^+$,

linear $\Leftrightarrow f(d) = ad + b \quad \forall d$ mit $a \in \mathbb{R}^+$ und $b \in \mathbb{R}$.

Name	Formel	Invarianz der Ergebnisse bei ¹			Skalen- transf.	Bemerkung
		Transla- tionen	nichtsing. lin. Tr.	orthogonale Rotation		
Varianz- kriterium	2.34	ja	nein	ja	nein	bei gleichgroßen Kugeln im Merkmalsraum
Gewichtete Varianzen	2.36	ja	nein	ja	nein	bei ungleich großen Kugeln im Merkmalsr.
Determinanten- kriterium	2.48	ja	ja	ja	ja	bei Ellipsoiden mit gleicher Achsrichtung; äquivalent zu WILKS- λ
Gewichtete Determinanten	2.50	ja	ja	ja	ja	bei verschiedenen Ellip- soiden; Bedingung: $ A_j > m \forall j$
HOTELLING- T^2 O	2.57	ja	ja	ja	ja	bei Ellipsoiden mit gleicher Achsrichtung
Größter Eigenwert	2.58	ja	ja	ja	ja	bei Ellipsoiden mit gleicher Achsrichtung
Abstandssum- menkrit.	2.59	ja	nein	ja	nein	-
Adaptive Distanzen	2.64	ja	ja	ja	ja	bei verschiedenen Ellip- soiden; Bedingung: $ A_j > m \forall j$ Optimallösung nicht notwendig konvex ² .

Tab 3.3: Eigenschaften von Zielfunktionen

- 1 Eine Transformation f der Zeilenvektoren der Datenmatrix heißt
Translation $\Leftrightarrow f(\mathbf{x}) = \mathbf{x} + \mathbf{b} \quad \forall \mathbf{x} \text{ mit } \mathbf{b} \in \mathbb{R}^m$,
nichtsinguläre lineare Transformation $\Leftrightarrow f(\mathbf{x}) = \mathbf{x} \mathbf{T} \quad \forall \mathbf{x} \text{ mit einer invertierbaren Matrix } \mathbf{T} \in \mathbb{R}^{(m,m)}$,
orthogonale Rotation $\Leftrightarrow f(\mathbf{x}) = \mathbf{x} \mathbf{O} \quad \forall \mathbf{x} \text{ mit einer orthogonalen Matrix } \mathbf{O} \in \mathbb{R}^{(m,m)}$,
Skalentransformation $\Leftrightarrow f(\mathbf{x}) = \mathbf{x} \mathbf{D} \quad \forall \mathbf{x} \text{ mit einer Diagonalmatrix } \mathbf{D} \in \mathbb{R}^{+(m,m)}$.
- 2 Eine Lösung heißt konvex, wenn die konvexen Hüllen der Cluster disjunkt sind. Die konvexe Hülle
des Clusters A_j ist die kleinste konvexe Teilmenge des $\mathbb{R}^m$, die alle Objekte des Clusters umfaßt.

Transformation	Formel	Eingabe	Ausgabe/Wirkung
Standardisierung (z-Transformation)	3.1	Datenmatrix	Mittelwert 0 und Varianz 1 bei allen Variablen
[0,1]-Normierung	3.4	Datenmatrix	Wertebereich [0,1] bei allen Variablen
Median-Normierung	3.5	Datenmatrix	Median 0 und Mittlere Absolute Abweichung 1 bei allen Variablen
Faktorenanalyse	-	Datenmatrix	Faktorenwerte von stochastisch unabhängigen Faktoren
Mehrdimensionale Skalierung	-	Distanzmatrix oder Ähnlichkeitsmatrix	Datenmatrix mit Koordinaten bezüglich künstlicher Raumachsen
Gewichtung	3.6	Datenmatrix	Skalentransformation zur Veränderung der Einflüsse der einzelnen Variablen
Umwandlung von Ähnlichkeitsmaßen in Distanzmaße	3.7	Ähnlichkeitsmatrix	Distanzmatrix
	3.8	Ähnlichkeitsmatrix mit $-1 \leq s_{ij} \leq 1$	Distanzmatrix
	3.9		

Tab. 3.4: Transformationen zur Datenaufbereitung

Name	Formel	Invarianz bei ¹				Bemerkung
		Transla- tionen	nichtsing. lin. Tr.	orthogonalen Rotationen	Skalen- transf.	
L-Metrik p (MINKOWSKI-Metrik)	3.10	ja	nein	nein	nein	-
City-Block- Distanz	3.10 mit $p=1$	ja	nein	nein	nein	-
Euklidische Distanz	3.10 mit $p=2$	ja	nein	ja	nein	-
Maximumsmetrik	3.11	ja	nein	nein	nein	-
Quadierte Eukli- dische Distanz	3.12	ja	nein	ja	nein	-
MAHALANOBIS- Distanz	3.13 mit $A=T^{-1}$	ja	ja	ja	ja	beseitigt Ungleich- gewichte und Kor- relationen
Produktmomentkor- relationskoeff.	3.14	ja	nein	nein	nein	vorherige Normie- rung der Variablen angebracht

Tab. 3.5: Distanz- und Ähnlichkeitsmaße für metrisch skalierte Merkmale

1 Vgl. Fußnote 1 auf S. 77.

Name	Formel	Skalenniveau der Merkmale	Wertebereich der Merkmale	Bemerkung
SPEARMAN'scher Rangkorrelations- koeffizient	3.16	rangskaliert	mehrwertig	-
Koeffizient von TANIMOTO	3.17 mit $\alpha=0$ und $\beta=0$	nominalskaliert	zweiwertig in $\{0,1\}$	gewichtet Ausprägung 1 stärker als Aus- prägung 0
Simple Matching Coefficient	3.17 mit $\alpha=1$ und $\beta=1$	nominalskaliert	zweiwertig in $\{0,1\}$	gewichtet beide Ausprägungen gleich
Verallgemeinerter Simple Matching Coefficient	3.19 bzw. 3.20	nominalskaliert	mehrwertig	gewichtet alle Ausprägungen gleich
Kombinierte Maße	3.22	gemischt	x	zusammengesetzt aus je einem Maß für je- des Skalenniveau

Tab. 3.6: Ähnlichkeitsmaße für nichtmetrische Merkmalsätze

In Abbildung 3.5 wird nun bei den Verarbeitungsschritten jeweils angegeben, welche der in diesen Tabellen angegebenen Bausteine in Frage kommen; zur endgültigen Entscheidung können die an einigen Stellen in den Tabellen unter "Bemerkung" bzw. einige im laufenden Text angegebenen Besonderheiten mit der spezifischen Anwendungssituation in Beziehung gesetzt werden. Normalerweise wird aber auch aufgrund solcher Zusatzinformationen keine eindeutige Auswahl getroffen werden können. Eine Entscheidung für eine bestimmte, d.h. von einem bestimmten Verfahren zu produzierende Lösung ist daher meistens ex ante nicht möglich; derzeit kann die weitere Einengung des Lösungsspektrums nur auf die ex-post-Betrachtung der verschiedenen von den zulässigen Verfahren erzeugten Lösungen verschoben werden. Die Chance, dabei "gute" oder "richtige" Ergebnisse auszuwählen, wird anhand von Simulationsexperimenten in dem nun folgenden Kapitel untersucht.

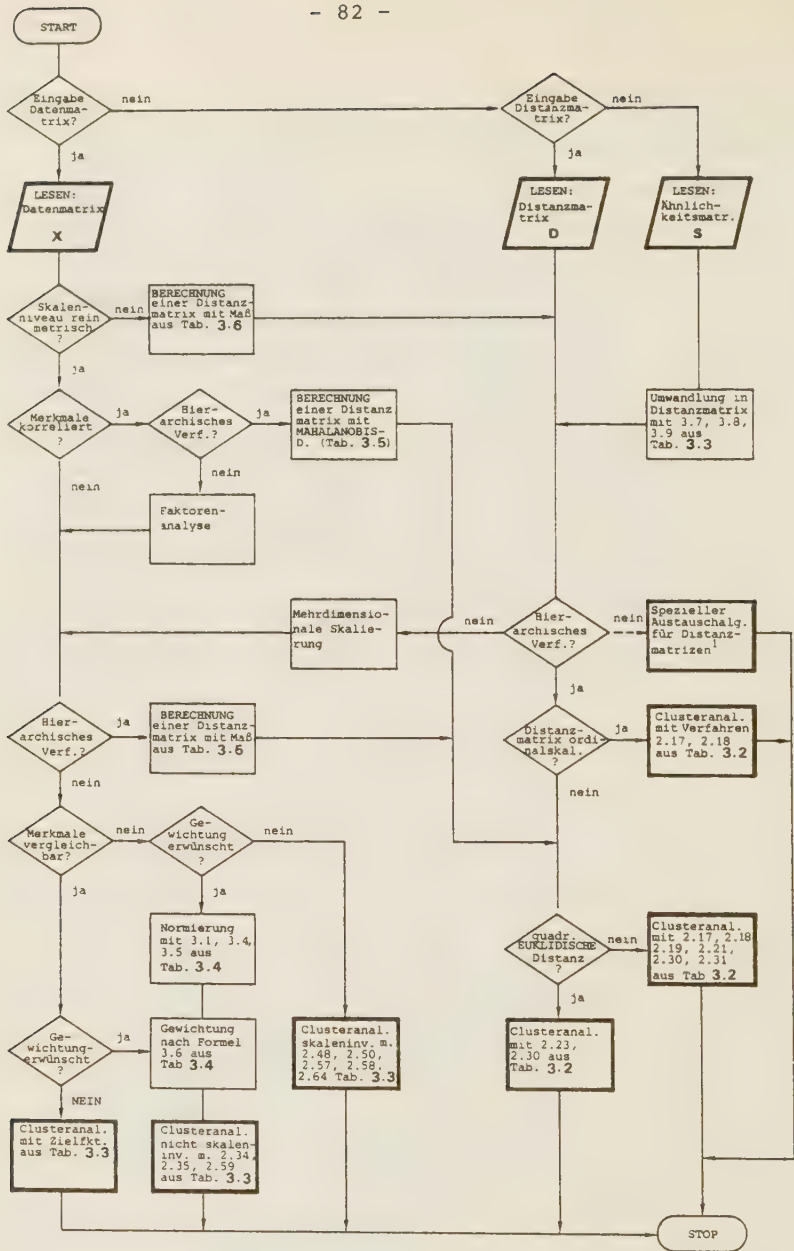


Abb. 3.5: Orientierungshilfe zur Kombination von methodischen Bausteinen bei Clusteranalysen

- 1 Austauschalgorithmen können auch zur Optimierung von Zielfunktionen benutzt werden, die direkt von einer Distanzmatrix **D** abhängen. Vgl. STEINHAUSEN, D., LANGER, K.: Clusteranalyse, Einführung in Methoden und Verfahren der automatischen Klassifikation. a.a.O., S. 135 ff.

4 DIE ANALYSE VON DATENSTRUKTUREN

4.1 SIMULATIONSEXPERIMENTE ZUR STRUKTURANALYSE

Eine der beiden Hauptaufgaben¹, für die Clusteranalysealgorithmen eingesetzt werden, ist das Aufdecken von Strukturen, die im Datensatz vorhanden sind und sich durch eine Gruppeneinteilung angemessen beschreiben lassen. Dazu ist es notwendig,

- zu erkennen, wieviele Cluster die zu analysierende Objektmenge enthält, und
- die bestmögliche Zuordnung zu treffen.

Da alle Verfahren "blind" ablaufen, d.h. auch dann Ergebnisse produzieren, wenn keine Cluster vorhanden sind, sollte darüberhinaus auch geprüft werden,

- ob im Datensatz überhaupt eine Gruppenstruktur vorhanden ist oder nicht.

Darin liegt aber zugleich das Kernproblem: Weder für den Begriff "Struktur" noch für "Cluster" gibt es eine erschöpfende Definition, die zugleich so konkret wäre, daß sie sich in mathematischen Formeln ausdrücken ließe.

Struktur als "Menge der Relationen, die die Elemente eines Systems miteinander verbinden"², ist ein vieldeutiger Ausdruck der Umgangssprache, der die verschiedensten Eigenschaften einer Objektmenge umfaßt. Auch eine Gruppenstruktur kann daher einmal durch die eine Eigenschaft, z.B. geringe Abstände der Objekte zum nächsten Nachbarn, ein anderes Mal durch eine andere Eigenschaft, wie etwa große Abstände der Clustermittelwerte, charakterisiert

¹ Vgl. dazu Abschnitt 1.2, S. 10 ff.

² WAHRIG, G.: Deutsches Wörterbuch. Gütersloh, Berlin, Neuauf-
lage 1977, Spalte 3579

werden; d.h. es gibt offenbar mehrere Strukturtypen, von denen jedes Verfahren nur einen erkennen kann. Die mit einer bestimmten Methode erzeugten Partitionen können deshalb nur dann eine tatsächlich vorhandene Datenstruktur wiedergeben, wenn diese von dem speziellen Typ ist, den das Verfahren von seiner Konzeption her erkennen kann. Solchen Objektmengen, die gar keine Struktur aufweisen oder eine andere, die möglicherweise mit einem anderen Verfahren sehr wohl zu erkennen wäre, wird zwangsläufig eine für das angewendete Verfahren, aber nicht für die Objektmenge charakteristische Gruppeneinteilung aufgeprägt, die die "richtigen" Cluster sogar verdecken kann.

Da "a-priori"-Informationen über den vorliegenden Strukturtyp i.d.R. aber nicht vorhanden sind, können die richtigen Verfahren, von den durch formale Voraussetzungen an die Daten bedingten Einschränkungen nach Abschnitt 3.3 einmal abgesehen¹, vor der Analyse nicht bestimmt werden. Weitere Informationen erhält man erst in Form der fertigen Ergebnisse, denen man aber wiederum nicht ansehen kann, ob sie reine Artefakte oder das Bild einer vorgefundenen Struktur sind. Eine objektive Auswahl kann deshalb nur dann getroffen werden, wenn außer der Partition, Hierarchie o.ä. mit den Ergebnissen ein Indikator geliefert wird, der anzeigt, ob das Verfahren seine spezifischen Voraussetzungen angetroffen hat oder nicht.

Solche Indikatoren muß man zwangsläufig in den Zielfunktionen bzw. Homogenitäts- und Heterogenitätsmaßen sehen, die zur Steuerung des Zuordnungsprozesses in den Algorithmen dienen; durch diese Verwendung wird ja gerade un-

1 vgl. S. 74 ff.

terstellt, daß ein besserer Funktionswert auch eine ausgeprägtere Struktur bedeutet, also als Maßzahl für eine bestimmte Form von "Struktur" zu verwenden ist. In diesem Fall müßte ein Datensatz, der tatsächlich eine solche Struktur aufweist, bei optimaler Clusterung auch einen besseren Funktionswert aufweisen als ein im übrigen gleicher, aber andersartig strukturierter oder gar nicht strukturierter Datensatz. Damit kommt den Zielfunktionen nicht nur die interne Aufgabe zu, die Zuordnung der einzelnen Objekte zu den Clustern zu steuern, sondern sie sollten gleichzeitig einen Anhaltspunkt bieten, anhand dessen beurteilt werden kann, ob im Datensatz eine solche Struktur vorgefunden wurde, die das Verfahren erkennen kann.

Mit vergleichenden Untersuchungen, bei denen Objektmengen mit bekannten, als richtig angesehenen Gruppeneinteilungen geclustert und die Ergebnisse den richtigen Clusterungen gegenübergestellt werden, kann beurteilt werden, mit welcher Methode unter bestimmten Voraussetzungen gute Gruppierungen erzielt werden können¹. Dieses Vorgehen kann besonders dann sinnvoll sein, wenn der Vergleich im Hinblick auf die Analyse eines bestimmten Datensatzes mit bekannten Charakteristika durchgeführt wird². Jedoch wird durch die Tatsache, daß ein Verfahren unter bestimmten Bedingungen gute Ergebnisse findet, noch nicht gewährleistet, daß diese Ergebnisse unter einer Reihe anderer Lösungen auch als gut erkannt werden. Insbesondere bei völlig fehlenden Vorinformationen über die mögliche Struktur der zu untersuchenden Objektmenge kann daraus kein Hinweis zur Auswahl der richtigen Methode, der richtigen Clusterzahl und der richtigen Partition abgeleitet werden.

1 vgl. z.B. KUIPER, K.F., FISHER, L.: A Monte Carlo Comparison of Six Clustering Procedures. *Biometrics* 31 (1975), S. 777-783.

2 so z.B. SITTERBERG, G.: Multivariate Analyse der Struktur und Entwicklung von Städten. Beiträge zum Siedlungs- und Wohnungswesen und zur Raumplanung Bd. 41, Münster 1977.

Im folgenden sollen nun verschiedene synthetische Datensätze mit mehreren Clusteranalyseverfahren bearbeitet werden, um zu untersuchen, inwiefern die mitgelieferten Gütemaße eine Entscheidung zwischen verschiedenen Clusterlösungen zulassen; dazu gehört natürlich auch die Frage, wie "richtig" die gefundenen Lösungen sind, denn so wenig es nützt, eine richtige Lösung zu produzieren, die nicht als solche erkannt wird, so wenig nützt es, eine eindeutige Entscheidung treffen zu können, wenn diese auf ein schlechtes Ergebnis hinausläuft.

Auch dieses Vorgehen kann - unvermeidlicherweise - unter zwei Gesichtspunkten kritisiert werden:

- Durch die Vorgabe von Datensätzen mit bestimmten Eigenschaften ist die Gültigkeit der Ergebnisse auf Datensätze mit diesen Eigenschaften beschränkt und läßt sich nicht ohne weiteres verallgemeinern.
- Die Richtigkeit eines Ergebnisses kann nur an der vorgegebenen Gruppenstruktur gemessen werden. Da "Struktur" aber nicht exakt definiert werden kann, ist möglicherweise bereits diejenige Gruppeneinteilung, die als die richtige vorgegeben wird, nicht eindeutig und damit als Maßstab untauglich.

Der erste Einwand läßt sich bei Simulationsexperimenten wohl nie ganz entkräften; man kann ihm jedoch begegnen, indem die Eigenschaften der synthetischen Datensätze möglichst allgemein gehalten werden, d.h. daß die Verteilungen der Zufallszahlen, aus denen die Daten erzeugt werden, in der Realität häufig vorkommen bzw. als Approximation akzeptabel sind. Dafür bieten sich Mischverteilungen aus mehrdimensionalen Normalverteilungen mit unterschiedlichen Parametern an. Der zweite Einwand kann nur dadurch ausgeräumt werden, daß die vorgegebenen, durch ein Clusteranalyseverfahren zu reproduzierenden Gruppenstrukturen sehr grob gewählt werden: Sie müssen so deutlich sein, daß sie auch bei unterschiedlichen Auffassungen von

"Clustern" und "Datenstrukturen" als solche erkannt und akzeptiert werden können. Damit entziehen sich natürlich die Grenzfälle schwer erkennbarer Cluster einer solchen Untersuchung. Ob ein Verfahren die nötigen Entscheidungshilfen bietet, wird sich jedoch bereits an einfachen, recht groben Kombinationen verschiedener Verteilungen erkennen lassen.

Die Datensätze wurden als Stichproben aus mehreren mehrdimensionalen Normalverteilungen $N_m(\mu, \Sigma)$ zusammengesetzt; diese wurden erzeugt, indem zunächst eindimensionale, stochastisch unabhängige, standardnormalverteilte Zufallsvariable $Y_1, \dots, Y_m$ mittels des SSP-Unterprogramms GAUSS¹ simuliert wurden. Zerlegt man nun die gewünschte Kovarianzmatrix Σ so, daß

$$\Sigma = A A^t \quad \text{mit} \quad A \in R^{(m,m)} \quad (4.1)$$

gilt, so genügt der aus $y = (Y_1, \dots, Y_m)$ durch die Transformation

$$x = y A + \mu \quad (4.2)$$

gebildete Zufallsvektor der gesuchten Verteilung²:

$$x \sim N_m(\mu, \Sigma) \quad (4.3)$$

1 SCIENTIFIC SUBROUTINE PACKAGE (SSP), Version III, Programmer's Manual. a.a.O., S. 77.

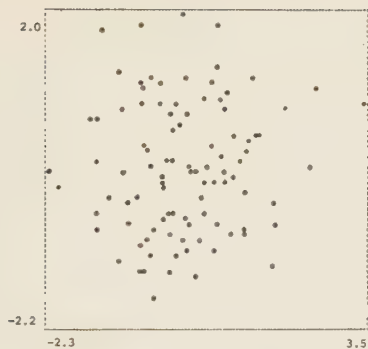
2 s. SCHMITZ, N., LEHMANN, F.: Monte-Carlo-Methoden I, Erzeugen und Testen von Zufallszahlen. Meisenheim am Glan 1976, S. 88-90.

Für die Zusammenstellung einer Verteilungsmischung wäre es naheliegend gewesen, auch die Parameter der beteiligten Verteilungen zufällig zu erzeugen. Wegen der erforderlichen deutlichen Trennung der Gruppen wurden die Parameter jedoch systematisch gewählt, weil ein Clusterprozeß¹, bei dem zunächst die Parameter zufällig erzeugt werden und in einer zweiten Stufe abhängig von diesen Parametern die Cluster generiert werden, keine deutlichen Gruppierungen gewährleisten kann, die als Maßstab für die Richtigkeit von Clusteranalyseergebnissen geeignet sind. Bei der Zusammenstellung der Mischverteilungen, wurden die folgenden Merkmale unabhängig voneinander variiert:

- die Anzahl der verschiedenen Verteilungen,
- die Gewichte der einzelnen Verteilungen,
- die Abstände der Mittelwertvektoren und
- die Kovarianzmatrizen.

Es gibt keinen Grund, anzunehmen, daß die Charakteristika der verschiedenen Methoden von einer höheren Dimensionszahl grundsätzlich beeinflußt werden; dagegen wird die Verarbeitung der Ergebnisse in zwei Dimensionen wesentlich dadurch erleichtert, daß sie leicht graphisch dargestellt werden können und visuell zu erfassen sind. Deshalb wurden für alle Datensätze zweidimensionale Verteilungen gewählt (s. Abb. 4.1).

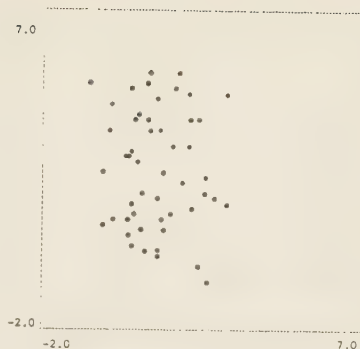
¹ Einige Beispiele bekannter Clusterprozesse mit Abbildungen finden sich bei STOYAN, D.: Applied Stochastic Geometry: A Survey. Biometrical Journal 21 (1979), S. 693-715.



Datensatz I (100 Objekte)

$$\mu = (0; 0) \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

100 Objekte

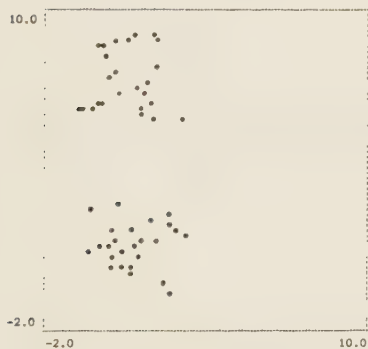


Datensatz II (50 Objekte)

$$\mu_1 = (1; 1) \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu_2 = (1; 4) \quad \Sigma_2 = \Sigma_1$$

jeweils 25 Objekte

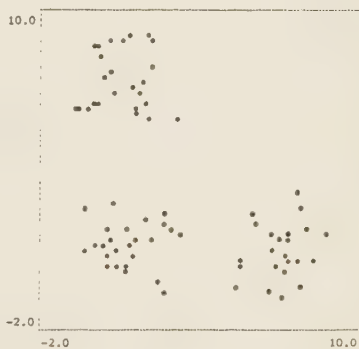


Datensatz III (50 Objekte)

$$\mu_1 = (1; 1) \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu_2 = (1; 7) \quad \Sigma_2 = \Sigma_1$$

jeweils 25 Objekte



Datensatz IV (75 Objekte)

$$\mu_1 = (1; 1) \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu_2 = (1; 7) \quad \Sigma_2 = \Sigma_1$$

$$\mu_3 = (7; 1) \quad \Sigma_3 = \Sigma_1$$

jeweils 25 Objekte

Abb. 4.1: Scatterdiagramme der verwendeten zweidimensionalen Zufallsdatensätze

16.0



Datensatz V (100 Objekte)

$$\mu_1 = (1; 1) \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

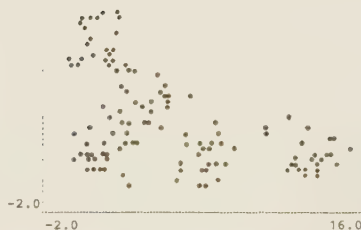
$$\mu_2 = (1; 7) \quad \Sigma_2 = \Sigma_1$$

$$\mu_3 = (7; 1) \quad \Sigma_3 = \Sigma_1$$

$$\mu_4 = (13; 1) \quad \Sigma_4 = \Sigma_1$$

jeweils 25 Objekte

16.0



Datensatz VI (125 Objekte)

$$\mu_1 = (1; 1) \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu_2 = (1; 7) \quad \Sigma_2 = \Sigma_1$$

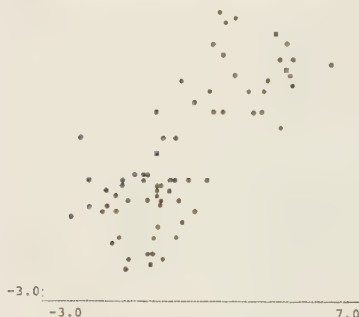
$$\mu_3 = (7; 1) \quad \Sigma_3 = \Sigma_1$$

$$\mu_4 = (13; 1) \quad \Sigma_4 = \Sigma_1$$

$$\mu_5 = (4; 4) \quad \Sigma_5 = \Sigma_1$$

jeweils 25 Objekte

7.0



Datensatz VII (75 Objekte)

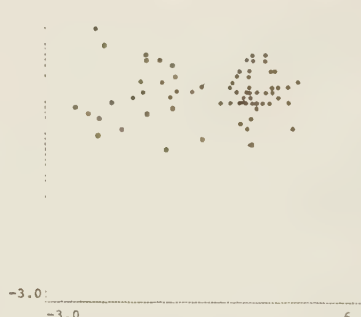
$$\mu_1 = (0; 0) \quad \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

50 Objekte

$$\mu_2 = (4; 4) \quad \Sigma_2 = \Sigma_1$$

25 Objekte

6.0



Datensatz VIII (75 Objekte)

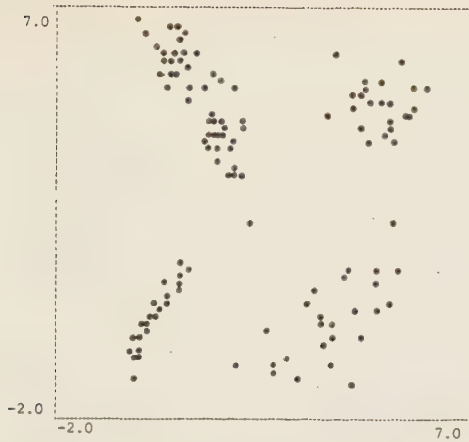
$$\mu_1 = (0; 0) \quad \Sigma_1 = \begin{pmatrix} 0,25 & 0 \\ 0 & 0,25 \end{pmatrix}$$

50 Objekte

$$\mu_2 = (2; 0) \quad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

25 Objekte

Abb. 4.1 (Fortsetzung): Scatterdiagramme der verwendeten zweidimensionalen Zufallsdatensätze



Datensatz IX (125 Objekte)

$$\mu_1 = (0; 0) \quad \Sigma_1 = \begin{pmatrix} 0,75 & 0,80 \\ 0,80 & 1,00 \end{pmatrix}$$

25 Objekte

$$\mu_2 = (4; 0) \quad \Sigma_2 = \begin{pmatrix} 1,48 & 0,24 \\ 0,24 & 1,44 \end{pmatrix}$$

25 Objekte

$$\mu_3 = (1; 5) \quad \Sigma_3 = \begin{pmatrix} 1,06 & -0,75 \\ -0,75 & 2,25 \end{pmatrix}$$

50 Objekte

$$\mu_4 = (5; 5) \quad \Sigma_4 = \begin{pmatrix} 0,25 & 0 \\ 0 & 0,25 \end{pmatrix}$$

25 Objekte

Abb. 4.1 (Fortsetzung): Scatterdiagramme der verwendeten zwei-dimensionalen Zufallsdatensätze

Das Spektrum verschiedener Mischungen wird noch dadurch erweitert, daß sich durch eine vorhergehende Normierung wie die z-Transformation (3.1) aus den bereits vorhandenen Datensätzen neue ergeben, indem z.B. vorher kugelförmige Cluster durch die unterschiedliche Streckung der Achsen zu Ellipsoiden werden (s. z.B. Abb. 4.2).

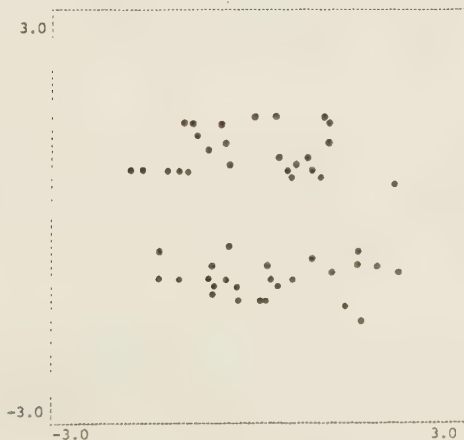


Abb. 4.2: Datensatz III nach z-Transformation

4.2 HEURISTISCHE REGELN ZUR AUSWAHL DER CLUSTERLÖSUNG

Die Auswahl der richtigen Clusterlösung wird häufig anhand der Entwicklung des Optimalitätskriteriums bzw. des Heterogenitätsmaßes¹ bei schrittweiser Änderung der Clusterzahl getroffen. Dahinter steht die Vorstellung, daß z.B. das Varianzkriterium bei zunehmender Clusterzahl auf jeden Fall kleiner wird, das absolute Minimum also i.A. bei der maximalen Clusterzahl erreicht wird und kein Anzeichen für die Zahl der Cluster im Datensatz ist; während jedoch die Werte beim Clustern einer strukturlosen Objektmenge gleichmäßig abnehmen müßten, sollten sie bei der Zerlegung einer Menge aus k Clustern zunächst stark, bei mehr als k Clustern aber nur noch geringfügig abnehmen, weil die weitere Verringerung ja nicht mehr auf vorhandene Gruppenstrukturen zurückzuführen ist². Trägt man nun in einem Diagramm den Wert des Kriteriums gegen die Clusterzahl ab (s. Abb. 4.3), so müßte die richtige Lösung an einem Knick (Elbow) im Kurvenverlauf erkennbar werden. Wenn sich kein deutlicher Knick ergibt, muß man annehmen, daß die untersuchte Objektmenge entweder gar keine Cluster enthält oder aus solchen Clustern besteht, die das gerade benutzte Verfahren nicht erkennen kann.

1 s.S. 31

2 vgl. auch SOEUR, W.: Empirische Verfahren zur Klassifikation. a.a.O., S. 125.

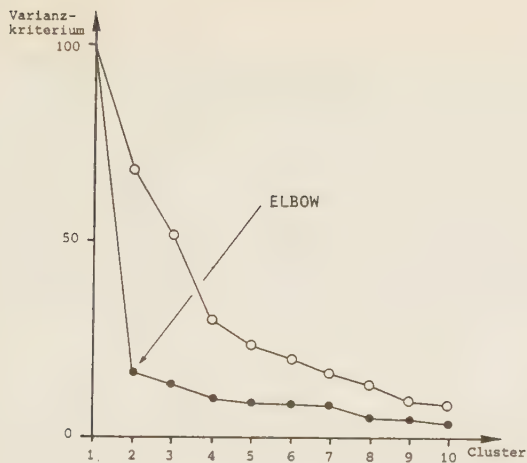


Abb. 4.3: Entwicklung des Varianzkriteriums in einer Objektmenge mit zwei Clustern (●-●-●) und einer Objektmenge ohne Cluster (O-O-O)

Um zu prüfen, inwiefern die mitgelieferten Gütemaße durch einen Elbow erkennen lassen, ob eine Struktur vorliegt und welche Clusterzahl die richtige ist, wurden die Datensätze II bis IX sowie III, VI, VIII und IX noch einmal nach einer vorangegangenen z-Transformation allen genannten Verfahren unterworfen¹. Für die hierarchischen Verfahren² (1) bis (7) sowie (8) mit $\alpha=0,6$ und $\alpha=0,7$ wurden quadrierte euklidische Distanzen (3.12) verwendet; beim Austauschalgorithmus wurden für die Kriterien $K_4(\alpha, X)$ und $K_8(\alpha, X)$ ³ Standardstartpartitionen⁴, für alle anderen Kriterien Zufallsstartpartitionen vorgegeben. Bereits dabei stellte sich heraus, daß die Kriterien $K_2(\alpha, X)$ und $K_4(\alpha, X)$ (s. (2.36) und (2.50)), die jeweils Potenzen und

1 Alle Rechenarbeiten wurden auf der IBM-3032 des Rechenzentrums der WWU in Münster durchgeführt.

2 s. Abschnitt 2.2

3 s. Abschnitt 2.3

4 In einer Standardstartpartition mit k Clustern wird zunächst Objekt i dem Cluster mit der Nummer $i \pmod k + 1$ zugeordnet.

Produkte von Varianzen und Determinanten enthalten, mit steigenden Objekt- und Clusterzahlen so kleine bzw. große Werte annahmen, daß die Berechnungen in den meisten Fällen wegen Überschreitung der rechentechnischen Genauigkeitsgrenzen abgebrochen werden mußten. Obwohl die Ergebnisse in den wenigen Läufen, die zuende geführt werden konnten, mit den Lösungen bei anderen Zielfunktionen vergleichbar waren, wurden diese Kriterien wegen der rechentechnischen Schwierigkeiten nicht weiter berücksichtigt.

Die mitgelieferten Gütemaße, das sind die Heterogenitätsmaße bei den hierarchischen Algorithmen und die Zielfunktionswerte bei den Austauschalgorithmen, wurden nun jeweils wie in Abb. 4.3 gegen die Clusterzahlen abgetragen und als Elbow wurde jeweils der schärfste Knick des Polygonzuges ungeachtet der absoluten Größe des Winkels angesprochen. Diese Elbows als Indikatoren für die vermeintlich richtige Gruppenzahl führten zu den in Tab. 4.1 angegebenen Zahlen; die richtig erkannten Clusterzahlen wurden darin noch besonders gekennzeichnet.

Verfahren	Datensatz	II	III	III*	IV	V	VI	VI*	VII	VIII	VIII*	IX	IX*	zutreffende Hinweise von Elbows
Hierarchische Verf.	Single Linkage	3	<u>2</u>	5	4	5	4	2	<u>2</u>	4	6	<u>4</u>	<u>4</u>	4
	Complete Linkage	4	<u>2</u>	2	4	<u>4</u>	<u>5</u>	3	3	3	4	2	<u>4</u>	5
	Average Linkage	3	<u>2</u>	3	4	<u>4</u>	<u>5</u>	<u>5</u>	<u>2</u>	<u>2</u>	5	3	7	6
	Weighted Av. Link.	<u>2</u>	<u>2</u>	4	4	<u>4</u>	<u>5</u>	3	5	5	8	<u>4</u>	<u>4</u>	6
	Median-Verf.	2	<u>2</u>	5	4	<u>4</u>	3	<u>5</u>	2	5	5	<u>4</u>	3	6
	Centroid-Verf.	6	<u>2</u>	6	4	<u>4</u>	<u>5</u>	<u>5</u>	<u>2</u>	6	8	2	<u>4</u>	6
	WARDS Verf.	<u>2</u>	<u>2</u>	4	<u>3</u>	<u>4</u>	<u>5</u>	<u>5</u>	2	<u>2</u>	6	<u>4</u>	<u>4</u>	10
	Flexible $\alpha=0.6$	4	2	5	<u>3</u>	<u>4</u>	<u>5</u>	3	2	3	6	2	<u>4</u>	6
	Strategie $\alpha=0.7$	<u>2</u>	<u>2</u>	5	<u>3</u>	<u>4</u>	2	3	<u>2</u>	<u>2</u>	4	2	<u>4</u>	7
Zielfunktionen	Varianzkriterium	<u>2</u>	<u>2</u>	4	<u>3</u>	7	2	3	<u>2</u>	<u>2</u>	4	<u>4</u>	<u>4</u>	7
	Abstandssummenk.	<u>2</u>	<u>2</u>	4	<u>3</u>	5	<u>5</u>	3	<u>2</u>	<u>2</u>	6	5	<u>4</u>	7
	WILKS' Lambda	<u>2</u>	3	3	<u>3</u>	2	3	3	<u>2</u>	<u>2</u>	7	2	3	4
	Größter Eigenwert	6	8	4	8	9	8	8	4	8	8	8	3	0
	HOTELLING- T_0^2	6	9	3	5	3	<u>5</u>	6	7	7	3	8	<u>4</u>	2
	Adaptive Distanzen	<u>2</u>	9	9	<u>3</u>	4	6	6	<u>2</u>	<u>2</u>	<u>2</u>	<u>4</u>	<u>4</u>	8

* Clusterung nach vorhergehender z-Transformation.

x stimmt mit der tatsächlichen Clusterzahl überein

Tab. 4.1: Durch Elbows hervorgehobene Clusterzahlen

Der Hinweis auf eine bestimmte Lösung ist natürlich nur dann verwertbar, wenn diese Lösung auch richtig ist, d.h. möglichst weitgehend mit der richtigen Partition übereinstimmt. Daher wurden alle Lösungen mit der richtigen Clusterzahl, auch wenn sie nicht von einem Elbow gekennzeichnet waren, daraufhin untersucht, wieviele Elemente den richtigen Clustern zugeordnet wurden¹. Diese Zahlen sind in Tab. 4.2 angegeben.

Verfahren	Datensatz	II	III	III*	IV	V	VI	VI*	VII	VIII	VIII*	IX	IX*	# richtig zugeordnete in %
Hierarchische Verf.	Single Linkage	27	50	50	75	99	-	71	49	52	52	125	125	92,16
	Complete Linkage	46	50	28	75	99	114	107	71	75	56	125	119	90,69
	Average Linkage	46	50	32	75	99	114	121	74	75	70	124	119	94,12
	Weighted Av. Link.	46	50	50	75	99	114	111	71	75	56	125	125	95,03
	Median-Verf.	46	50	32	75	99	105	100	71	52	52	100	76	82,76
	Centroid-Verf.	46	50	50	75	99	114	115	71	75	52	124	124	94,12
	WARDs Verf.	45	50	49	75	99	114	114	74	75	73	125	125	97,12
	Flexible $\alpha=0.6$ Strategie $\alpha=0.7$	46	50	47	75	99	114	120	74	75	70	125	125	97,02
Zielfunktionen		46	50	49	75	99	114	112	71	75	70	125	125	96,82
	Varianzkriterium	46	50	50	75	69	121	121	75	69	70	122	123	94,66
	Abstandssummenkrit.	46	50	50	75	71	121	121	75	66	67	124	124	94,36
	WILKS' Lambda	26	30	46	75	99	122	108	75	67	64	123	101	86,74
	Größter Eigenwert	39	50	34	60	60	64	112	75	67	65	68	73	80,81
	HOTELLING- T_0^2	41	30	45	70	-	103	97	75	67	64	77	92	81,38
	Adaptive Distanzen	46	50	50	75	99	92	92	75	67	67	125	78	89,94

* Clusterung nach vorhergehender z-Transformation

Tab. 4.2: Anzahl richtig zugeordneter Elemente

1 Dabei wurde immer die günstigste Permutation der Cluster zugrundegelegt, d.h. es wurde so gezählt, daß sich die größtmögliche Zahl richtiger Zuordnungen ergab.

Selbst bei den hier benutzten, recht groben Testdatensätzen wird deutlich, daß die von den Eigenwerten von $\mathbf{W}^{-1}\mathbf{B}$ abhängenden Kriterien $K_3'(\mathbf{0},\mathbf{X})$, $K_5(\mathbf{0},\mathbf{X})$ und $K_6(\mathbf{0},\mathbf{X})$ (s. (2.56), (2.57) und (2.58)) offenbar nicht anzeigen können, ob eine gute oder schlechte Lösung gefunden wurde. Obwohl in einigen Fällen vollständig richtige oder zumindest fast vollständig richtige Lösungen gefunden wurden, wird dies aus dem Verlauf der Zielfunktionswerte in keinem Fall erkennbar. Das Single-Linkage Verfahren wiederum weist mit größtenteils sehr ausgeprägten Elbows meist auf die falsche Clusterzahl hin, auch wenn das Dendrogramm unter anderem die richtige Lösung enthält. Das Median-Verfahren liefert gegenüber den restlichen Verfahren auffällig schwache Zuordnungen.

Auf der anderen Seite liefert das Verfahren von WARD mit durchschnittlich 97,12 % richtig zugeordneten Elementen sehr gute Lösungen, die gleichzeitig in 10 von 12 Fällen am Heterogenitätsmaß auch richtig erkannt werden können. Obwohl die Zuordnungen bei der Flexiblen Strategie und den drei verbleibenden Zielfunktionen $K_1(\mathbf{0},\mathbf{X})$, $K_7(\mathbf{0},\mathbf{X})$ und $K_8(\mathbf{0},\mathbf{X})$ im Austauschalgorithmus nicht nennenswert schlechter sind, bietet das Verfahren von WARD den Vorteil, daß eine im Datensatz enthaltene Gruppenstruktur besser gekennzeichnet wird. Zudem lassen die Dendrogramme, die die hierarchischen Verfahren erzeugen, eventuelle Verzerrungen durch Ausreißer, die als eigenes Cluster angesehen werden, besser erkennen als die Ergebnislisten der Austauschalgorithmen, die lediglich aufzählen, welches Objekt zu welchem Cluster gehört.

Zusammenfassend hat sich bei diesem Experiment ergeben, daß

- die Kriterien $K_2(\mathbf{0},\mathbf{X})$ und $K_4(\mathbf{0},\mathbf{X})$ wegen rechentechnischer Schwierigkeiten als Zielfunktion untauglich sind;

- die Eigenwertkriterien $K'_3(\mathbf{A}, \mathbf{X})$, $K_5(\mathbf{A}, \mathbf{X})$ und $K_6(\mathbf{A}, \mathbf{X})$ die richtige Clusterzahl nicht erkennen lassen;
- der Elbow beim Single Linkage-Verfahren mehr falsche als richtige Hinweise gibt und das Median-Verfahren schlechte Zuordnungen findet;
- nur das Verfahren von WARD gleichzeitig sehr gute Partitionen findet und meistens die richtige Clusterzahl signalisiert.

Leider werden selbst beim Verfahren nach WARD auch sehr gute Partitionen nicht mit Sicherheit durch Elbows angezeigt und es kann auch nicht ausgeschlossen werden, daß eine unzutreffende Partition von einem Elbow hervorgehoben wird. Noch nicht einmal durch die Übereinstimmung der Elbows bei mehreren Verfahren kann die Trefferquote erhöht werden. Deshalb wäre ein sichereres Auswahlkriterium wünschenswert.

Da die beobachteten Kriterienwerte einerseits von der Datenstruktur, andererseits von der Güte der Clusterung beeinflusst werden, ist es naheliegend, die Wirkungen dieser beiden Einflußgrößen zu trennen, indem

- die suboptimalen Kriterienwerte, die von Clusteranalyseverfahren erzielt werden, mit den Werten verglichen werden, die man erhält, wenn derselbe Datensatz willkürlich oder zufallsgesteuert geclustert wird. Dadurch wird der Effekt des systematischen Clusters isoliert.
- Andererseits können die suboptimalen Werte mit den Werten verglichen werden, die sich ergeben, wenn ein vergleichbarer Datensatz (d.h. gleich groß, gleiche Varianzen etc.), der aber bekanntermaßen unstrukturiert ist, möglichst gut geclustert wird. Dadurch wird der Einfluß der Datenstruktur isoliert¹.

1 vgl. auch BERGONIER, H., MORIN, B.: Typology at Random. ESOMAR Seminar on Segmentation and Typology, Brüssel 1972.

Versuche zur Zweckmäßigkeit beider Vorgehensweisen wurden wegen der Bedeutung, die der Varianz als Homogenitätsmaß in theoretischen und empirischen Arbeiten immer wieder zugemessen wird und die auch durch die Erfolge des Verfahrens von WARD und des Varianzkriteriums als Zielfunktion (s. Tab. 4.1 und 4.2) bestätigt wird, auf das Varianzkriterium beschränkt.

Bei der Gegenüberstellung der suboptimalen Kriteriumswerte aus dem Austauschalgorithmus und der Zufallswerte muß bedacht werden, daß auch eine Zufallsclustering besonders gut oder besonders schlecht sein kann. Deshalb müssen zur Stabilisierung der Vergleichswerte jeweils mehrere Zufallsergebnisse herangezogen werden, indem z.B. das arithmetische Mittel für den Vergleich benutzt wird. Im Experiment wurden deshalb zu den Datensätzen II bis IX zu jeder Clusterzahl zwischen 2 und 10 jeweils 100 Zufallslösungen erzeugt, zu denen anschließend der Wert des Varianzkriteriums berechnet wurde. Aus der empirischen Verteilung der Zufallswerte wurde dann jeweils der Mittelwert und dessen Differenz zum Kriteriumswert der bereits bekannten suboptimalen Lösung des Austauschalgorithmus bestimmt (s. Tab. 4.3).

Cluster	Sub-optimum	Mittelwert	Differenz	5%-Punkt	Differenz
1	250,00	-	-	-	-
2	112,26	247,96	135,70	244,63	132,37
3	54,93	245,60	190,67	240,85	185,92
4	39,55	244,45	204,90	239,59	200,04
5	21,75	242,27	220,52	234,98	213,23
6	18,98	239,97	220,99	230,39	211,41
7	16,20	238,42	222,22	229,39	213,19
8	14,35	235,97	221,62	225,66	211,31
9	10,78	232,98	222,20	221,71	210,93
10	11,20	233,04	221,84	221,77	210,57

maximale Differenz

Tab. 4.3: Abweichungen des Varianzkriteriums zwischen Suboptimum und Mittelwert bzw. 5%-Punkt der Empirischen Verteilung der Zufallswerte bei Datensatz VI nach einer z-Transformation.

Es zeigte sich, daß die Verteilung der Zufallswerte nicht erkennbar von der Struktur des Datensatzes abhängt; ob dieser aus einem oder mehreren Clustern besteht, der Mittelwert der Verteilung nimmt immer fast linear ab zwischen der Gesamtvarianz bei einem Cluster und 0 bei n Clustern (= Objektzahl) (s. Abb. 4.4), sodaß die Mittelwerte auch gut durch eine lineare Funktion approximiert werden können, die nur von n , m und der Gesamtvarianz abhängt.

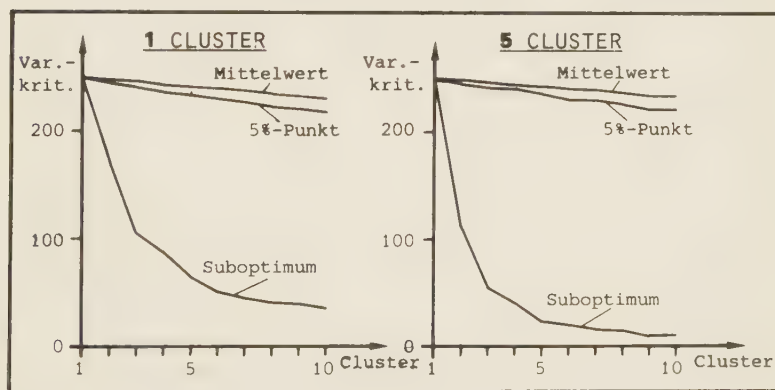


Abb. 4.4: Suboptimale Werte des Varianzkriteriums sowie Mittelwert und 5%-Punkt der empirischen Verteilung der Zufallswerte bei Datensätzen mit 125 Objekten und 1 bzw. 5 Clustern

Leider stellte sich heraus, daß die Differenz zwischen dem Kriteriumswert bei suboptimaler Clusterung und dem Mittelwert bei Zufallsclusterung ein schlechter Indikator für die Richtigkeit der Lösung ist; bei den hier verwendeten Datensätzen lag die maximale Differenz nur ein einziges Mal bei der richtigen Clusterzahl (s. Tab. 4.4), obwohl in der Reihe der suboptimalen Lösungen sehr gute Lösungen

zur richtigen Clusterzahl enthalten waren (vgl. Tab. 4.2). Dieses Ergebnis konnte auch durch die Verwendung des 5%-Punktes der empirischen Verteilung, der möglicherweise ein besserer Maßstab für eine gute Lösung ist als der Mittelwert, nicht nennenswert verbessert werden. Damit wurde die richtige Clusterzahl dreimal getroffen (s. Abb. 4.4 und Tab. 4.4).

Datensatz	Clusterzahl		
	aus Differenz zum Mittelwert	aus Differenz zum 5%-Punkt	richtig
II	7	5	2
III	4	2	2
III*	7	4	2
IV	5	4	3
V	7	7	4
VI	5	5	5
VI*	7	5	5
VII	8	6	2
VIII	9	6	2
VIII*	9	9	2
IX	5	5	4
IX*	5	5	4
richtige Hinweise	1	3	-

* Verarbeitung nach vorhergehender z-Transformation
☐ richtige Clusterzahl

Tab. 4.4: Clusterzahlen mit der maximalen Differenz zwischen Mittelwert bzw. 5%-Punkt der empirischen Verteilung der Zufallswerte und dem suboptimalen Wert

Das zweite Konzept ist die Isolation der Struktur des untersuchten Datensatzes, indem diesem ein unstrukturierter Datensatz gegenübergestellt wird, der aber gleichfalls möglichst gut im Sinne des verwendeten Optimalitätskriteriums geclustert wird. Die Differenz der Kriteriumswerte kann dann nur auf Unterschiede in der Beschaffenheit der beiden Datensätze zurückzuführen sein. Um die Vergleichbarkeit zu gewährleisten, müssen die Vergleichsdatsätze

hinsichtlich äußerlicher Eigenschaften wie Objektzahl, Variablenzahl u.ä. mit dem zu analysierenden Datensatz so weit wie möglich übereinstimmen, damit nicht von Eigenschaften, die nicht Gegenstand der Untersuchung sind, Wirkungen auf die Werte der Optimalitätskriterien ausgeübt werden.

Die andererseits geforderte Strukturlosigkeit des Vergleichsdatensatzes ist jedoch genauso wenig exakt definierbar, wie "Struktur" selbst: müssen die Objekte dazu völlig gleichmäßig, d.h. gesetzmäßig im Merkmalsraum verteilt sein, genügen Zufallsvektoren aus einer Gleichverteilung oder ist eine andere Wahrscheinlichkeitsverteilung richtiger? Die gleichen Gründe, die es sinnvoll erscheinen ließen, ein Cluster als Stichprobe aus einer mehrdimensionalen Normalverteilung und Gruppenstrukturen durch Mischungen verschiedener Verteilungen darzustellen¹, sprechen nun auch dafür, einen strukturlosen Datensatz als ein einziges Cluster anzusehen, also durch eine Stichprobe aus nur einer mehrdimensionalen Normalverteilung darzustellen.

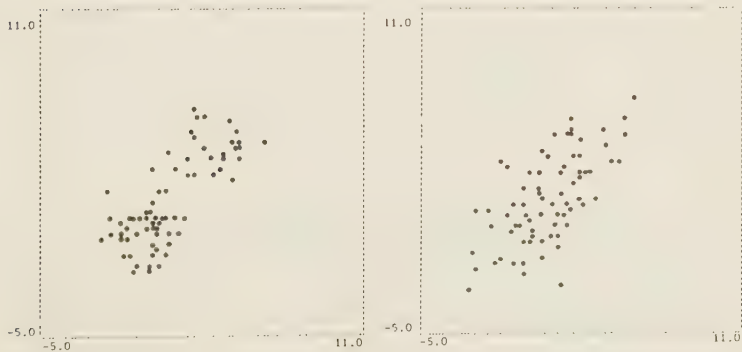


Abb. 4.5: *Datensatz VII und Vergleichsdatensatz mit einem Cluster*

1 s. S. 86

Wäre der untersuchte Datensatz nun unstrukturiert und aus derselben Verteilung $N_m(\mu, \Sigma)$ wie der Vergleichsdatsatz, so würden sich die Werte der Optimalitätskriterien beim Clustern nur geringfügig unterscheiden. Eine durch das Kriterium meßbare Datenstruktur müßte jedoch zu einer stärkeren Abweichung vom Vergleichsdatsatz führen, anhand derer man die richtige Clusterzahl und Partition herausfinden kann.

Als Beispiel zum Vergleich sind in Abb. 4.6 die Werte des Varianzkriteriums bei Datensatz VI mit 5 Clustern und einem Vergleichsdatsatz mit nur einem Cluster eingezeichnet.

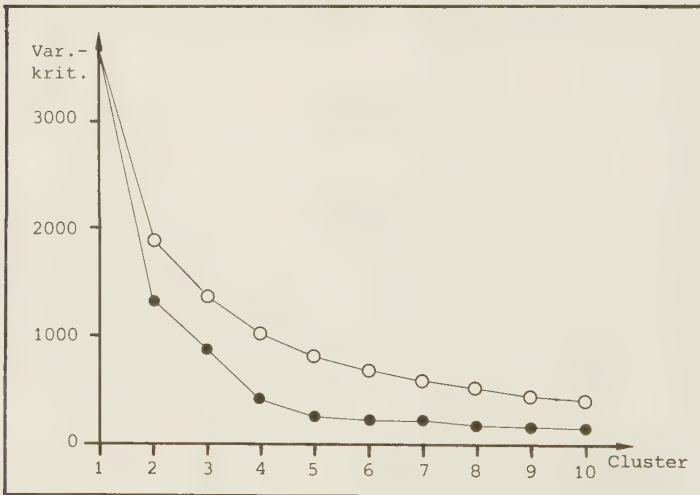


Abb. 4.6: Varianzkriterium bei Datensatz VI (●-●-●) und einem Vergleichsdatsatz (○-○-○)

Die Parameter μ und Σ der Verteilung des Vergleichsdatsatzes lassen sich zunächst über

$$\hat{\mu} = \bar{x} \quad (4.4)$$

und

$$\hat{\Sigma} = \frac{1}{n-1} T \quad (4.5)$$

erwartungstreu¹ aus dem Mittelwertvektor $\bar{x}$ und der Streu-

1 s. z.B. MORRISON, D.F.: Multivariate Statistical Methods. New York u.a., 1967, S. 97

ungsmatrix $\mathbf{T}$ des untersuchten Datensatzes schätzen und die Produktion von Zufallsvektoren, die der Verteilung $N_m(\hat{\mu}, \hat{\Sigma})$ genügen, verläuft genau wie bei der Erstellung der Testdatensätze¹; für die dabei notwendige Zerlegung von $\hat{\Sigma}$ in ein Matrizenprodukt $\mathbf{A}\mathbf{A}^t$ (s. (4.1)) kann ein bekanntes Verfahren von CHOLSKY benutzt werden, das i.d.R. aus Programmbibliotheken ohne nähere Kenntnis der technischen Einzelheiten abgerufen werden kann². Die zufallsbedingten Abweichungen der Streuungsmatrizen bei den so erzeugten Vergleichsdatsätzen waren im Experiment zunächst noch so groß, daß die Tragfähigkeit des Vergleichs darunter erheblich litt. Deshalb wurden zusätzlich alle Variablen linear so transformiert, daß sie dieselbe Stichprobenstandardabweichung wie die entsprechende Variable des untersuchten Datensatzes aufwiesen. Wegen der für $\alpha = \{I\}$ leicht zu zeigenden Beziehung

$$K_1(\alpha, X) = \sum_{j=1}^k \sum_{i \in A_j} \|x_i - \bar{x}_j\|^2 = (n-1) \sum_{j=1}^k s_j^2 \quad (4.6)$$

und der Übereinstimmung der Stichprobenstandardabweichungen s_j^2 aller Variablen bei beiden Datensätzen ist damit sichergestellt, daß die Werte des Varianzkriteriums bei einem Cluster bei den zu vergleichenden Datensätzen übereinstimmen. Daher müssen Unterschiede der Kriteriumswerte bei höheren Clusterzahlen auf Unterschiede der inneren Struktur der Datensätze zurückgeführt werden. Ein Beispiel eines Vergleichsdatsatzes, der wie oben beschrieben konstruiert wurde, zeigt die Abb. 4.5.

Auch bei dem hier angestrebten Vergleich können zufällig mehr oder weniger gut zu clusternde Zufallsdatensätze die Aussagekraft des Vergleichs noch empfindlich stören; deshalb sollten in jedem Falle die Ergebnisse von mehreren Vergleichsdatsätzen herangezogen werden. Dazu wurde von den Vergleichswerten, deren Zahl wegen der hohen Re-

1 s.S. 87

2 Hier wurde das Unterprogramm MFSD benutzt aus: SCIENTIFIC SUBROUTINE PACKAGE (SSP), Version III, Programmer's Manual, a.a.O., S. 158.

chenzeiten des Austauschalgorithmus auf jeweils drei beschränkt werden mußte, das arithmetische Mittel gebildet. Die größte Differenz dieses Mittelwertes zum Kriteriumswert des untersuchten Datensatzes wurde schließlich als Hinweis auf die günstigste Clusterzahl respektive die beste Partition angesehen (s.Abb. 4.6). Welche Clusterzahlen sich auf diese Weise bei den Testdatensätzen II bis IX als die besten herausstellten, ist zusammen mit den richtigen Clusterzahlen in Tab. 4.5 aufgeführt.

Datensatz	Clusterzahl	
	aus Vergleich	richtig
II	2!	2
III	2!	2
III*	4	2
IV	3!	3
V	4	4
VI	4?	5
VI*	3?	5
VII	2!	2
VIII	2	2
VIII*	5?	2
IX	3?	4
IX*	4	4
richtige Hinweise	7	-

* Verarbeitung nach vorhergehender z-Transformation

☐ richtige Clusterzahl

! maximale Differenz sehr groß

? maximale Differenz sehr klein

Tab. 4.5: Clusterzahlen mit der maximalen Differenz zwischen dem Kriteriumswert beim untersuchten Datensatz und bei den umstrukturierten Vergleichsdatsätzen

Die Anzahl richtig erkannter Lösungen ist bei diesem Vorgehen nicht größer als beim Elbow-Kriterium (vgl. Tab. 4.1); berücksichtigt man jedoch nicht nur die Rangfolge der Differenzen, sondern zusätzlich auch deren Größenverhältnisse, so ergibt sich, daß

- besonders deutliche Unterschiede zwischen der größten und der nächstkleineren Differenz, d.h. hier ein Unterschied von mehr als 1/3 der größeren Differenz, stets die richtige Clusterzahlangaben (in Tab. 4.5 mit ! gekennzeichnet),
- geringe Unterschiede zwischen der größten und nächstkleineren Differenz, d.h. hier weniger als 10 % der größeren, stets eine falsche Clusterzahl ergaben (in Tab. 4.5 mit ? gekennzeichnet).

Dadurch wird zwar die Zahl der deutlichen Hinweise auf 4 gesenkt, immerhin bekommt man dadurch aber einen Maßstab für die Zuverlässigkeit dieses Auswahlkriteriums, der beim Elbow-Kriterium so leicht nicht konstruiert werden konnte.

Eine ähnliche Methode zur Auswahl der richtigen Clusterlösung ist z.B. von BERGONIER¹ entwickelt worden und findet in dem von INFRATEST benutzten Typologieprogramm SYAT² Anwendung. Dort wird für jede Variable 1, $1 \leq l \leq m$, der Quotient

$$R2_1 = \frac{\sum_{j=1}^k n_j (\bar{x}_{j1} - \bar{x}_1)^2}{\sum_{j=1}^k \sum_{i \in A_j} (x_{i1} - \bar{x}_1)^2} \cdot 1000 \quad (4.7)$$

als der durch die Typologie nicht erklärte Varianzanteil definiert und für den gesamten Datensatz der durchschnittliche R2-Wert definiert als

$$\overline{R2} = \frac{1}{m} \sum_{l=1}^m R2_l \quad (4.8)$$

1 vgl. BERGONIER, H., MORIN, B.: Typology at Random. a.a.O., S.5 ff.

2 s. STADTLER, K.: Zur Methodik des Infratest-Typologie-Programms SYAT. München 1975.

Dem Wert von R^2 (der bei standardisierten Variablen eine lineare Funktion des Varianzkriteriums ist) wird nun der Wert RH^2 gegenübergestellt; das ist der Wert von R^2 , der sich ergeben würde, wenn "alle Variablenwerte zufällig verteilt wären"¹, womit eine nicht näher bezeichnete Normalverteilung gemeint ist². Da diese Zufallsverteilung offenbar vom untersuchten Datensatz nur über die Objekt- und Variablenzahl abhängt, kann für RH^2 auch eine nur von n und m abhängige Näherungsformel angegeben werden. Auch in diesem System wird die maximale Differenz zwischen R^2 und RH^2 als Anzeichen für die optimale Typenzahl angesehen. Dies wird jedoch nur plausibel begründet; eine empirische Untersuchung, inwiefern diese Differenz in der Lage ist, eine vorhandene Clusterstruktur und die Anzahl der Cluster richtig anzuzeigen, steht noch aus³.

1 s. STADTLER, K.: Zur Methodik des Infratest-Typologie-Programms SYAT. a.a.O., S. 19.

2 s. BERGONIER, H., MORIN, B.: Typology at Random. a.a.O., S. 5.

3 Der Vergleich mit Datensätzen, die nur durch n und m mit dem untersuchten Datensatz zusammenhängen, nicht aber die Gestalt der Punktwolke im Merkmalsraum z.B. durch Varianzen und Kovarianzen berücksichtigten, brachte im zuvor beschriebenen Experiment deutlich schlechtere Ergebnisse.

4.3 TESTS AUF SIGNIFIKANZ DER CLUSTERLÖSUNG

Mit den heuristischen Auswahlregeln des letzten Abschnitts kann immer nur diejenige aus einer Reihe von Lösungen ausgewählt werden, die sich durch einen relativ guten Kriteriumswert, einen Elbow usw. am deutlichsten heraushebt und von der man deshalb annimmt, daß sie die beste der vorliegenden Lösungen darstellt. Damit wird aber leider die Möglichkeit nicht ausgeschlossen, daß auch diese nur die beste aus einer Serie von schlechten Lösungen und damit selbst auch schlecht ist. Nur die letzte der geschilderten Methoden, nämlich der Vergleich mit einem Datensatz, der nur ein Cluster enthält, gibt einen vagen Hinweis darauf, ob die gefundene Lösung ein Artefakt wie alle anderen ist oder sich dadurch vor allen anderen auszeichnet, daß der Gruppierungsvorgang durch eine vorhandene Datenstruktur erleichtert wird und so zu einem besonders guten Wert des Optimalitätskriteriums führt. Dieser Hinweis ist jedoch nur in extremen Ausprägungen verläßlich und stellt in anderen Fällen keine Hilfe mehr dar.

Da einige der zum Clustern verwendeten Optimalitätskriterien ohnehin der Testtheorie entlehnte Prüfgrößen für Tests auf Mittelwertunterschiede sind, liegt der Gedanke nahe, mit Hilfe dieser Prüfgrößen Signifikanzaussagen derart zu machen, daß

- beim Unter- bzw. Überschreiten eines kritischen Wertes vorbehaltlich eines Irrtumsniveaus auf das Vorhandensein mehrerer Cluster geschlossen werden kann und
- im anderen Falle im Wert der Prüfgröße keine Anzeichen für mehrere Cluster im Datensatz gesehen werden können.

Tatsächlich sind die Nullverteilungen, d.i. die Verteilung der Prüfgröße unter der Annahme der Nullhypothese, z.B. des WILKS'schen Λ (2.56), des HOTELLING- T_0^2 (2.57) und des Größten Eigenwertes (2.58) vom Testen Linearer Hypothesen

her bekannt und vertafelt¹. Die Prozentpunkte dieser Verteilungen können jedoch nicht als kritische Werte für einen Test auf die Signifikanz von Clusterstrukturen verwendet werden², weil zur Herleitung dieser Verteilungen auf jeden Fall die stochastische Unabhängigkeit aller Einzelbeobachtungen vorausgesetzt wird. Durch das Clustern vor der Berechnung des Kriteriums mit dem Ziel der systematischen Veränderung der Prüfgröße wird jedoch diese Annahme zunichte gemacht, sodaß auch unstrukturierte Datensätze nach dem Clustern fast ausschließlich Werte aufweisen, die im Test Linearer Hypothesen hochsignifikant sind. Vergleicht man z.B. die empirischen Verteilungen des Varianzkriteriums (2.34), die sich ergeben, wenn man (vgl. Abb. 4.7)

- Datensätze, die als Stichproben aus *einer* mehrdimensionalen Normalverteilung gewonnen werden, zufällig clustert, was der Nullhypothese beim Testen auf Mittelwertunterschiede entspricht (Verteilung I in Abb. 4.7), oder
- dieselben Datensätze mit dem Ziel, das Varianzkriterium zu minimieren, clustert (Verteilung II in Abb. 4.7), oder
- Datensätze, die als Stichproben aus einer *Mischverteilung* gewonnen wurden (Verteilung III in Abb. 4.7), optimal clustert,

so erkennt man, daß die beiden letzten fast vollständig links vom 5%-Punkt der ersten liegen³; d.h. egal ob der geclusterte Datensatz strukturiert ist (Vtlg. III) oder nicht (Vtlg. II), der bestmögliche Wert des Varianzkriteriums liegt auf jeden Fall im kritischen Bereich eines entsprechenden Tests auf Mittelwertunterschiede, sodaß der kritische Wert kaum noch Aussagekraft hat.

1 S. z.B. die umfassende Sammlung von KRES, H.: Statistische Tafeln zur multivariaten Analysis. Berlin, Heidelberg, New York 1975.

2 Vgl. auch EVERITT, B.S.: Unresolved Problems in Cluster Analysis. a.a.O., S. 171.

3 Die Vergleichbarkeit von Datensätzen mit einem und mit mehreren Clustern wurde durch eine Normierung der Gesamtstreuung hergestellt.
- Zur Ermittlung der Verteilung optimaler Werte s. S. 114 f.

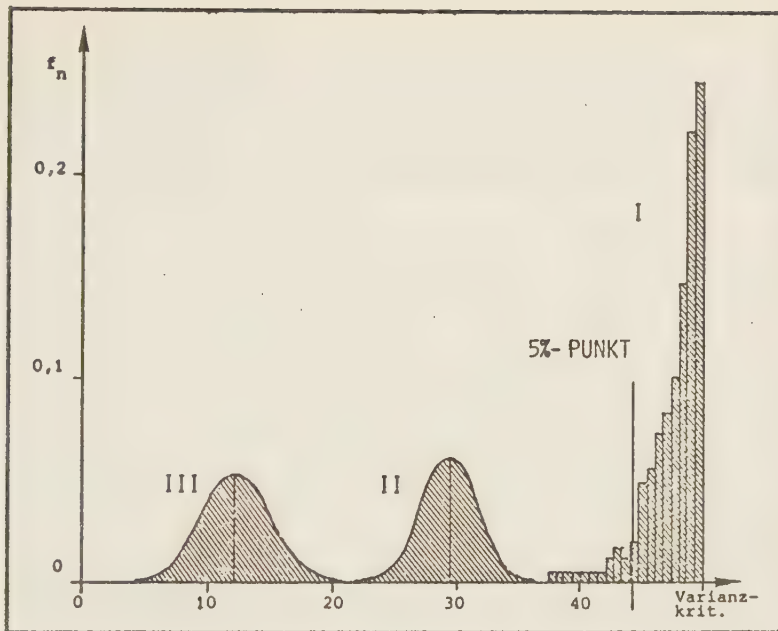


Abb. 4.7: Empirische Verteilung des Varianzkriteriums bei Zufallsclustering (I) und optimaler Clustering (II) unstrukturierter Datensätze und bei optimaler Clustering (III) strukturierter Datensätze

Für einen Test zur Unterscheidung zwischen strukturierten und unstrukturierten Datensätzen würden vielmehr die Prozentpunkte der Verteilung II, d.h. die Verteilung von

$$\min\{K_1(\alpha, X) / \alpha \in \mathcal{P}(I), |\alpha| = k\} \quad (4.9)$$

bei einem zu minimierenden Kriterium i bzw. von

$$\max\{K_1(\alpha, X) / \alpha \in \mathcal{P}(I), |\alpha| = k\} \quad (4.10)$$

bei einem zu maximierenden Kriterium i und einer Clusterzahl k benötigt. Diese Verteilungen der bestmöglichen Kriteriumswerte, die sich an unstrukturierten Datensätzen erzielen lassen, sind aber leider in keinem Fall bekannt. Auch die Nichtnullverteilungen, also die Verteilungen der Prüfgröße unter der Annahme der Gegenhypothese, deren

Kenntnis für die Konstruktion eines statistischen Tests ebenfalls wichtig wäre, sind völlig unbekannt, zum größten Teil schon deshalb, weil es noch nicht einmal gelungen ist, präzise Gegenhypothesen zu formulieren.

Aus den bekannten Verteilungen der Prüfgrößen können also keinesfalls Signifikanzaussagen, sondern lediglich Hinweise auf die richtige Lösung abgeleitet werden, indem - wie im Beispiel in Abb. 4.8 - die durch Clustern erzielten Werte der Prüfgröße mit den Prozentpunkten der Verteilung verglichen werden. Die größte Differenz kann nun wieder als Hinweis auf die günstigste Partition angesehen werden. Dieses Vorgehen ist aber theoretisch nicht besser fundiert als die Heuristiken des letzten Abschnittes und entspricht vom Konzept und der Treffsicherheit her dem Vergleich mit dem Mittelwert bzw. den Prozentpunkten der empirischen Verteilung der Kriteriumswerte bei zufälliger Clusterung¹.

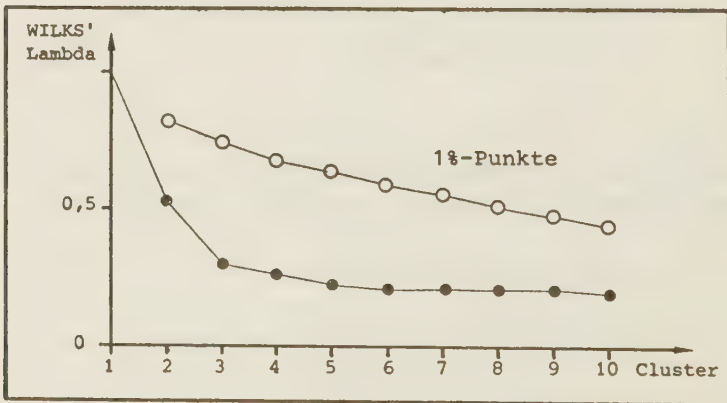


Abb. 4.8: Untere 1%-Punkte der Nullverteilung des WILKS'schen Lambda (O—O—O) und Werte des WILKS'schen Lambda bei Datensatz II (●—●—●)²

1 vgl. Abschnitt 4.2 auf S. 92

2 Die Prozentpunkte sind KRES, H.: Statistische Tafeln zur multivariaten Analysis. a.a.O., Tafel 1, S. 20 ff entnommen. Die fehlenden Werte wurden linear interpoliert.

Etwas bessere Ergebnisse lieferten dagegen die beiden folgenden Kriterien¹, die sich bereits bekannte Verteilungen von Varianzenquotienten zunutze machen. Bezeichnet man mit $K_1(i)$ den besten erreichbaren Wert des Varianzkriteriums bei i Clustern, also

$$K_1(i) = \min\{K_1(\alpha, X) / \alpha \in \mathcal{P}(I), |\alpha| = i\} \quad , \quad (4.11)$$

dann soll mit

$$K_1(i, j) = \frac{K_1(i) - K_1(j)}{K_1(j)} \bigg/ \left(\frac{n-1}{n-j} \left(\frac{j}{i} \right)^{2/m} - 1 \right) \quad (4.12)$$

die Signifikanz der Abnahme des Varianzkriteriums beim Übergang von i zu j Clustern mit $i < j$ beurteilt werden: Bei einem Wert von $K_1(i, j)$, der kleiner als der kritische Wert F_{α, n_1, n_2} der F-Verteilung mit $n_1 = m(j-i)$ und $n_2 = m(n-j)$ Freiheitsgraden ist, besteht zwischen der i - und der j -Cluster-Lösung "kein Unterschied"², der auf besondere Qualitäten einer der beiden Lösungen zurückgeführt werden könnte. Dieser Fall ist jedoch bei allen Testdatensätzen kaum jemals vorgekommen. Betrachtet man dagegen nur die Werte $K_1(i, i+1)$ für $i = 1, \dots, n-2$, so kann der größte dieser Werte unabhängig von irgendeiner vermuteten Verteilung als Hinweis auf die richtige Lösung angesehen werden. Wie aus Tab. 4.6 erkennbar wird, konnten damit in 7 von 12 Fällen die richtigen Lösungen bestimmt werden.

Das gleiche gilt auch für den aus der Varianzanalyse bekannten Quotienten

$$K'_1(i, j) = \frac{K_1(i) - K_1(j)}{K_1(j)} \frac{n-j}{j-1} \quad (4.13)$$

für $j = 2, \dots, n-1$,

der im Falle $m = 2$ ohnehin zu $K_1(1, j)$ proportional ist. Auch hier konnten anhand des größten Wertes 6 von 12 Lösungen richtig erkannt werden.

1 Vgl. z.B. BÜTTNER, J.: Zur Clusteranalyse. Biometrische Zeitschrift 17 (1975), S. 163 - 179

2 Ebenda, S. 174

Datensatz	Clusterzahlen		
	aus $K_1(i, i+1)$	aus $K_1^1(1, j)$	richtig
II	2	2	2
III	2	2	2
III*	4	4	2
IV	3	3	3
V	7	7	4
VI	4	5	5
VI*	5	9	5
VII	2	2	2
VIII	9	2	2
VIII*	4	8	2
IX	4	5	4
IX*	4	5	4
richtige Hinweise	7	6	-

* Verarbeitung nach vorhergehender z-Transformation
☐ richtige Clusterzahl

Tab. 4.6: Clusterzahlen mit dem größten Wert von $K_1(i, i+1)$ bzw. $K_1^1(1, j)$

Ein Test der Clusterstruktur aufgrund der Zielfunktionswerte scheint zur Zeit nicht konstruierbar zu sein, weil die Stichprobenverteilung der optimalen Zielfunktionswerte analytisch nicht bestimmt werden kann. Von BRYANT¹ sind jedoch einige Tests angegeben worden, die es erlauben, unter bestimmten Voraussetzungen anhand der Matrix T bzw. ihrer Eigenwerte zwischen den Hypothesen

$$H_0: \mathbf{x}_1 \sim N_m(\boldsymbol{\mu}, \Sigma) \quad i, 1 \leq i \leq n \quad (4.14)$$

und

$$H_A: \mathbf{x}_1 \sim \sum_{j=1}^k g_j N_m(\boldsymbol{\mu}_j, \Sigma) \quad i, 1 \leq i \leq n \quad (4.15)$$

$$\text{mit } \sum_{j=1}^k g_j = 1 \quad \text{und} \quad \Sigma = \sigma^2 I_m, \quad \sigma^2 \in \mathbb{R}^+,$$

also zwischen einer einfachen Normalverteilung und einer Mischverteilung mit k Komponenten zu entscheiden. Dieser Test hängt nur von T ab und hat damit den Vorteil, daß er

1 BRYANT, P.: On Testing for Clusters Using the Sample Covariance. Journal of Multivariate Analysis 5 (1975) S. 96-105.

unbeeinflusst vom zufälligen Erfolg eines Clusteralgorithmus eine Aussage darüber zuläßt, ob die Objektmenge Cluster aufweist, und wenn ja, wieviele. Gleichzeitig hat er aber auch den Nachteil, daß er damit zwar eine bestimmte Clusterzahl auszeichnet, aber nicht eine bestimmte Partition, sodaß die Auswahl der Ergebnisse zwar eingeschränkt, aber nicht endgültig gelöst werden kann. Außerdem sind diese Tests bislang auf einfache Situationen beschränkt, bei denen aus grundsätzlichen Erwägungen der für die Clusteranalyse letztlich wichtigste Fall, daß Σ beliebig und unbekannt ist, ausgeschlossen werden muß. Deshalb ist auch aus dieser Richtung für realistische Clusterprobleme keine wesentliche Hilfe zu erwarten.

Der Versuch, einen Test für Clusteranalyseergebnisse auf der Basis des Optimalitätskriteriums analytisch abzuleiten, scheint derzeit aussichtslos zu sein. In solchen Fällen bietet sich mit Monte-Carlo-Methoden noch eine Möglichkeit an, zumindest Teilergebnisse zu erkunden. Die Untersuchung der hier benötigten Testverteilungen wird jedoch i.A. dadurch verhindert, daß es keine Algorithmen gibt, die das globale Optimum irgendeines Optimalitätskriteriums innerhalb vernünftiger Bearbeitungszeiten mit Sicherheit finden. Um die Verteilungen von (4.9) bzw. (4.10) erkennen zu können, würden außerdem nicht nur einzelne, sondern jeweils eine große Anzahl von optimalen Ergebnissen benötigt, was beim derzeitigen Stande der verfügbaren Algorithmen undurchführbar ist.

Eine Ausnahme bildet lediglich das Varianzkriterium. Dadurch, daß gezeigt werden konnte, daß Partitionen, die bezüglich des Varianzkriteriums optimal sind, konvex¹ sein müssen, wird die Zahl der in Betracht kommenden Clusterlösungen wesentlich geringer als $S(n,k)$ bzw. $B(n)^2$, sodaß das Optimum leichter gefunden werden kann. Insbesondere im Spezialfall

1 Eine Partition $\mathcal{U} = \{A_1, \dots, A_k\}$ heißt konvex, wenn die konvexen Hüllen der Cluster $A_1, \dots, A_k$ disjunkt sind. S. auch BOCK, H.H.: Automatische Klassifikation. a.a.O., S. 170, Satz 15.3.

2 vgl. Tab. 2.2 und 2.3 auf S. 45

m=1 und k=2 (1 Variable und 2 Cluster) wird die Optimierungsaufgabe auf die Bestimmung eines Schnittpunktes reduziert, der die beiden rechts und links von diesem Schnittpunkt gelegenen Cluster trennt. Diese Vereinfachung macht die vollständige Enumeration sämtlicher konvexer Lösungen, die in anderen Fällen immer noch sehr aufwendig ist, möglich. Damit konnten ENGELMAN und HARTIGAN¹ bereits 1969 die oberen Prozentpunkte der Nullverteilung von

$$\frac{K_1(1) - K_1(2)}{K_1(2)} \quad (4.16)$$

angeben, die sie aus je 100 000 Stichproben n normalverteilter Zufallszahlen mit $3 \leq n \leq 50$ ermittelten.

Obwohl inzwischen Algorithmen zur Verfügung stehen, die das globale Optimum des Varianzkriteriums i.d.R. ohne Totalenumeration finden, können die Stichprobenverteilungen für andere Fälle als den oben genannten, also größere Variablen- und Clusterzahlen, immer noch nicht mit den gleichen Genauigkeitsanforderungen untersucht werden, weil die benötigten Rechenzeiten immer noch extrem hoch sind. So muß beim Clustern von 25 Objekten mit 2 Variablen in 6 Cluster mit dem Unterprogramm BBOUND² damit gerechnet werden, daß die CPU-Time auf der IBM 3032 des Rechenzentrums der Universität Münster 15 Minuten übersteigt. An ähnlich große Stichprobenumfänge, d.h. so viele Optimallösungen wie bei ENGELMAN und HARTIGAN ist daher keinesfalls zu denken.

Deshalb wurden die folgenden Versuche mit jeweils 25 Objekten und 2 Variablen nur mit jeweils 21 Testdatensätzen durchgeführt. Das Ziel des Vorhabens war es, festzustellen, ob die

1 ENGELMAN, L., HARTIGAN, J.A.: Percentage Points of a Test for Clusters. Journal of the American Statistical Association 64 (1969), S. 1647-1648

2 S. STEINHAUSEN, D., LANGER, K.: Ein global optimaler Clusteranalysealgorithmus zur Optimierung von Zielfunktionen. a.a.O.

Nullhypothese: Die Stichprobe entstammt einer Normalverteilung,
gegen die

Alternativhypothese: Die Stichprobe entstammt einer Mischung aus mehreren Normalverteilungen,

getestet werden kann. Die Nullhypothese wurde dementsprechend durch zweidimensionale Zufallsvektoren $\mathbf{x}_i$ mit

$$\begin{aligned} \mathbf{x}_i &\sim N_2(\boldsymbol{\mu}, \Sigma) \quad \text{für } i = 1, \dots, 25 \\ \text{mit } \boldsymbol{\mu} &= (0,0) \quad \text{und } \Sigma = I_2 \end{aligned} \quad (4.17)$$

und eine mögliche Alternativhypothese durch

$$\begin{aligned} \mathbf{x}_i &\sim \begin{cases} N_2(\boldsymbol{\mu}_1, \Sigma) & \text{für } i = 1, \dots, 13 \\ N_2(\boldsymbol{\mu}_2, \Sigma) & \text{für } i = 14, \dots, 25 \end{cases} \\ \text{mit } \boldsymbol{\mu}_1 &= (0,0), \quad \boldsymbol{\mu}_2 = (3,3) \quad \text{und } \Sigma = I_2 \end{aligned} \quad (4.18)$$

simuliert. Sowohl zur Null- als auch zur Gegenhypothese wurden jeweils 21 Datensätze mit Hilfe der Subroutine GAUSS¹ gebildet, die dann mit der Subroutine BBOUND in 2 bis 5 Cluster zerlegt wurden. Für die Untersuchung der Verteilung von $K_1(i)$ (s. (4.11)) standen also unter der Annahme der Null- wie der Gegenhypothese zu den Clusterzahlen $i = 2, 3, 4, 5$ jeweils 21 Werte zur Verfügung.

Bei der weiteren Betrachtung stellte es sich als zweckmäßig heraus, statt $K_1(j)$ besser den Quotienten

$$K_1(j) / K_1(1) \quad \text{für } j = 2, \dots, 5 \quad (4.19)$$

zu betrachten, dessen Stichprobenmittelwerte und -standardabweichungen in Tab. 4.7 angegeben werden.

1 Scientific Subroutine Package (SSP), Version III, Programmer's Manual, a.a.O., S. 77.

	unstrukturiert				strukturiert			
Cluster	2	3	4	5	2	3	4	5
$\bar{x}$	0,554	0,338	0,229	0,166	0,246	0,162	0,116	0,085
s	0,046	0,037	0,029	0,020	0,053	0,029	0,024	0,018
d	0,084	0,124	0,142	0,086	0,116	0,174	0,188	0,143
$d_c (\alpha=0.05)$	0,190				0,190			

Tab. 4.7: Parameter der empirischen Verteilung von $K_1(j)/K_1(1)$ bei strukturierten und homogenen Datensätzen sowie Werte der KOLMOGOROV-SMIRNOV-Prüfgröße D bei Anpassung einer Normalverteilung und der kritische Wert d_c^1 .

Daraus ergeben sich nun folgende Informationen:

- Die Werte von $K_1(j)/K_1(1)$ sind bei den strukturierten Datensätzen deutlich niedriger als bei den unstrukturierten.
- Die Standardabweichungen liegen in beiden Fällen nahe beieinander und werden im F-Test nicht als signifikant ($\alpha=0,05$) verschieden ausgewiesen.

Obwohl dies günstige Voraussetzungen für einen Test sind, können die kritischen Werte aus der Häufigkeitsverteilung von nur 21 verschiedenen Werten nicht mit akzeptabler Genauigkeit ermittelt werden. Aus diesem Grunde und auch zur theoretischen Absicherung der Testkonstruktion ist außer den Stichprobenparametern $\bar{x}$ und s noch die Kenntnis des Verteilungstyps erforderlich.

1 Der hier verwendete Wert d_c wurde der Tabelle von LILLIEFORS entnommen, die speziell für das Testen von Normalverteilungen mit geschätzten Parametern durch Monte-Carlo-Methoden ermittelt wurde. S. LILLIEFORS, H.W.: On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association 62 (1967), S. 399 - 402.

In Anbetracht des geringen Stichprobenumfangs wurde der Kolmogorov-Smirnov-Test¹ herangezogen, um die Verträglichkeit der je 21 Werte mit verschiedenen häufig gebrauchten Verteilungen wie β -, Γ - und F-Verteilungen zu prüfen. Dabei ergab sich, daß bei gleichzeitiger Betrachtung aller Stichproben lediglich die Hypothese einer Normalverteilung in keinem Falle abgelehnt werden mußte (vgl. die Werte der Prüfgröße D in Tab. 4.7), obwohl eine Normalverteilung wegen

$$0 \leq K_1(j)/K_1(1) \leq 1 \quad (4.20)$$

grundsätzlich nur eine Näherung sein kann. Ist aber $K_1(j)/K_1(1)$ unter Null- und Gegenhypothese (wenigstens annähernd) normalverteilt und zwar mit der gleichen Varianz σ^2 , so wird durch die Entscheidungsregel

$$\begin{array}{ll} \text{Annahme} & \text{der } H_0 \text{ bei } K_1(j)/K_1(1) \geq k_c \\ \text{Ablehnung} & K_1(j)/K_1(1) < k_c \end{array} \quad (4.21)$$

mit $F_n(k_c/\mu_0; \sigma^2) = \alpha$ und μ_0 als Mittelwert der Nullverteilung

ein gleichmäßig-bester Test zum Niveau α formuliert²; der kritische Wert k_c dieses Tests kann also allein aus Mittelwert und Standardabweichung der Nullverteilung bestimmt werden.

Die o.a. Tabelle von Mittelwerten und Standardabweichungen ist sehr klein und eine Erweiterung mit Hilfe der Subroutine BBOUND wäre wegen der extremen Rechenzeiten kaum möglich. Bei den für die Tab. 4.7 durchgeführten Berechnungen

1 vgl. BLEYMÜLLER, J., GEHLERT, G., GÜLICHER, H.: Statistik für Wirtschaftswissenschaftler. a.a.O., S. 129 ff.

2 Zur Konstruktion gleichmäßig-bester (uniform most-powerful) Tests vgl. LEHMANN, E.L.: Testing Statistical Hypotheses. New York 1959, S. 60 ff.

stellte sich jedoch gleichzeitig heraus, daß auch der Austauschalgorithmus bei kleineren Clusterzahlen (2 bis 4) in mehr als zwei Drittel aller Fälle und bei den größeren Clusterzahlen (hier bis 6) immer noch in einem Drittel aller Fälle das globale Minimum des Varianzkriteriums findet. Gleicht man dies durch jeweils mehrere Läufe mit verschiedenen Startpartitionen aus, von denen man das beste Ergebnis verwendet, so findet man mit wesentlich geringerem Aufwand an Rechenzeit Ergebnisse, die kaum vom Optimum abweichen. Auf diese Weise konnten die folgenden Tabellen mit den Mittelwerten und Standardabweichungen der Nullverteilungen von $K_1(j)/K_1(1)$ ermittelt werden, die sich auf das Spektrum von 25 bis 125 Objekten mit 2 Variablen erstrecken, die in 2 bis 10 Cluster unterteilt werden.

Objekte Cluster	25	50	75	100	125
1	1,0	1,0	1,0	1,0	1,0
2	0,614	0,627	0,617	0,648	0,663
3	0,357	0,400	0,408	0,426	0,429
4	0,263	0,318	0,329	0,347	0,349
5	0,200	0,244	0,255	0,267	0,277
6	0,155	0,186	0,207	0,221	0,216
7	0,115	0,152	0,173	0,186	0,188
8	0,089	0,131	0,147	0,163	0,162
9	0,071	0,114	0,128	0,141	0,142
10	0,061	0,099	0,110	0,124	0,128

Tab. 4.8: Mittelwerte der Nullverteilungen

Objekte Cluster	25	50	75	100	125
1	0	0	0	0	0
2	0,049	0,024	0,070	0,026	0,027
3	0,028	0,021	0,018	0,018	0,019
4	0,032	0,027	0,024	0,026	0,024
5	0,033	0,023	0,019	0,010	0,015
6	0,026	0,020	0,017	0,021	0,015
7	0,016	0,019	0,014	0,015	0,015
8	0,011	0,018	0,014	0,013	0,011
9	0,012	0,016	0,009	0,008	0,010
10	0,013	0,010	0,009	0,009	0,009

Tab. 4.9: Standardabweichungen der Nullverteilungen

Mit Hilfe dieser Werte und der Entscheidungsregel (4.21) können nun Clusterlösungen auf signifikante Abweichungen von der Nullhypothese (4.17) getestet werden.

Faßt man nun diejenigen Lösungen, die sich vom Mittelwert μ der betreffenden Verteilung gemessen in Einheiten der Standardabweichung σ am stärksten unterscheiden, für die also

$$\frac{\mu - K_1(j) / K_1(1)}{\sigma} \quad (4.22)$$

maximal wird, als die besten auf, so bekommt man bei den bereits zuvor verwendeten Testdatensätzen die in Tab. 4.10 angegebenen Hinweise.

Datensatz	Clusterzahlen		
	nach (4.22)	aus Faustr.	richtig
II	☐ 2	☐ 2	2
III	☐ 2	☐ 2	2
III*	4	4	2
IV	☐ 3	☐ 3	3
V	5	4	4
VI	☐ 5	5	5
VI*	☐ 5	5	5
VII	3	2	2
VIII	3	2	2
VIII*	☐ 2	6	2
IX	5	6	4
IX*	6	☐ 4	4
richtige Hinweise	6	9	-

* Verarbeitung nach vorhergehender z-Transformation
☐ richtige Clusterzahl

Tab. 4.10: Clusterzahlen mit der stärksten Abweichung vom Mittelwert der Nullverteilung

Aufgrund der geringen Stichprobenumfänge, die den Tabellen 4.8 und 4.9 zugrunde liegen, sind diese Hinweise noch relativ ungenau. Setzt man aber statt der dort angegebenen Werte zur Glättung als grobe Faustregel z.B. $\mu = 1/j$ und $\sigma = 0,1 \cdot \mu$ in (4.22) ein, verwendet also

$$10 - 10 j \cdot K_1(j) / K_1(1) \quad (4.23)$$

anstelle von (4.22) als Maßstab, so bekommt man die ebenfalls in Tab. 4.10 angegebenen Hinweise, die bereits erheblich besser sind.

4.4 ZUSAMMENFASSUNG

Die Leistungsfähigkeit eines Clusteranalysealgorithmus bzw. eines sog. Optimalitätskriteriums als Hilfsmittel bei der Aufdeckung von Gruppenstrukturen in Datenmengen hängt nicht nur davon ab, ob mit der jeweiligen Methode die "richtige" Partition gefunden wird bzw. wie nahe man damit an die richtigen Lösungen herankommt. Gute Leistungen in dieser Hinsicht sind nur dann von Nutzen, wenn die richtige Lösung unter einer Fülle anderer Lösungen auch erkennbar wird.

Die experimentelle Prüfung anhand einiger Testdatensätze zeigte,

- daß einige der in der Literatur immer wieder genannten Optimalitätskriterien schon aus rechentechnischen Gründen ungeeignet sind;
- daß einige Verfahren von vornherein nur weniger gute Partitionen lieferten, die nicht an der Datenstruktur orientiert waren (vgl. Tab. 4.2) und
- daß bei den restlichen Verfahren durch einen Elbow in kaum mehr als 50 % der Fälle die richtige Lösung identifiziert wurde, wobei jedoch das hierarchische Verfahren von WARD, das implizit das Varianzkriterium minimiert, eine mit Abstand führende Stellung einnahm.

Die weiterhin nur noch mit dem Varianzkriterium durchgeführten Versuche ergaben,

- daß der Vergleich mit den Werten des Optimalitätskriteriums bei Zufallsclustering desselben Datensatzes kaum hilfreich ist und
- daß beim Vergleich mit einem unstrukturierten Zufallsdatensatz, der aber möglichst gut geclustert wird, etwa so viele richtige Lösungen gefunden wurden, wie beim Elbowkriterium. Diese Methode kann aber durch bessere Anpassung des Vergleichsdatsatzes vermutlich noch verbessert werden.

Verlangt man außer der Hervorhebung der besten Lösung noch einen Maßstab, der gestattet, zu beurteilen, ob eine Lösung eine tatsächliche Datenstruktur wiedergibt oder nur künstlich aufgeprägt ist, so findet man,

- daß das Elbow-Kriterium sehr in die Irre führen kann,
- daß der Vergleich mit Zufallsdatensätzen schwache Anhaltspunkte geben kann und
- daß mit den aus der Testtheorie abgeleiteten Prüfgrößen und Testverteilungen keine Signifikanzaussagen gemacht werden können. Als Auswahlkriterium für die beste Lösung sind sie nicht besser, als die bereits genannten heuristischen Methoden. Hinsichtlich der Möglichkeiten des Testens ergab sich,
- daß lediglich beim Varianzkriterium für eng begrenzte Spezialfälle die Nullverteilung eines Tests für Clusterstrukturen durch Monte-Carlo-Methoden approximiert werden konnte, sodaß in diesen Fällen Signifikanzaussagen möglich sind.

Die Erweiterung der Testmöglichkeiten scheint demnach die erfolgsversprechendste Möglichkeit zu sein, objektive Kriterien zur Beurteilung von Clusterergebnissen zu entwickeln. Diese würde nicht nur die Ausweitung der zugrundeliegenden Stichprobenumfänge und die Verbreiterung der erfaßten Bereiche für Objekt-, Variablen- und Clusterzahlen umfassen, sondern außerdem die Präzisierung anderer denkbarer Null- bzw. Gegenhypothese; bei anderen Testmodellen wiederum könnten andere Optimalitätskriterien als das hier vornehmlich untersuchte Varianzkriterium wirksamer sein.

5 DIE KLASSIFIKATION VON OBJEKTMENGEN

5.1 GÜTEMAßE FÜR KLASSIFIKATIONSVERFAHREN

In vielen praktischen Fällen wird mit der Gliederung einer Objektmenge nicht das Ziel verfolgt, darin enthaltene Cluster aufzudecken, sondern das Ziel, die Objektmenge, auch wenn sie nur aus einem einzigen großen Cluster im Sinne der Strukturanalyse¹ besteht, in irgendeinem Sinne zweckmäßig zu unterteilen. Dabei soll der Fall, daß die Zweckmäßigkeit einer solchen Unterteilung durch ein von vornherein definiertes Kriterium gemessen werden kann, wie z.B. die Kostenfunktion im Multiplen Standortproblem², hier nicht betrachtet werden, weil sich das Problem, die Güte einer Unterteilung zu messen bzw. das beste Ergebnis auszuwählen, dabei nicht stellt.

Häufig können jedoch derartige Kriterien nicht hergeleitet werden, weil z.B. das angestrebte Ziel der Klassifikation noch nicht bekannt ist oder nicht als reelle Funktion dargestellt werden kann. Beispielsweise werden in einer Datenbank häufig nicht formatierte Informationen abgespeichert mit dem Ziel, sie möglichst sicher bzw. mit möglichst geringem Recherche-Aufwand wiederfinden zu können. Da aber nicht bekannt ist, unter welchem speziellen Aspekt, d.h. anhand welcher Merkmale eine bestimmte Information später einmal gesucht werden wird, ist es kaum möglich, den späteren Suchaufwand und damit die Zweckmäßigkeit der Speichersystematik zu bemessen.

Ohne einen objektiven Maßstab bleibt aber die Auswahl eines geeigneten Klassifikationssystems zwangsläufig der subjektiven Wertschätzung eines menschlichen Betrachters überlassen, was aus mehreren bereits genannten Gründen³ unbefriedigend ist. Einerseits sind geringe Ka-

1 Vgl. S. 14

2 S. S. 11

3 Vgl. S. 3

pazitäten betreffs Objekt- und Variablenzahl gerade der Anlaß, mechanisierte Systeme zu benutzen, sodaß dem Betrachter auch nur eine beschränkte Urteilsfähigkeit bei größeren Datenmengen zugesprochen werden kann. Andererseits erscheinen in vielen Situationen mehrere Lösungen plausibel, sodaß auch Plausibilität nicht als objektives Auswahlkriterium angesehen werden kann.

Auch wenn es unter diesen Voraussetzungen offenbar nicht möglich ist, eine hinreichende Bedingung für Optimalität einer Klassifikation anzugeben, bedeutet dies noch nicht, daß die Auswahl von Lösungen gänzlich willkürlich geschehen müßte. Es können nämlich möglicherweise immer noch einige notwendige Bedingungen für die Optimalität angegeben werden, die zwar nicht ausreichen, um eine Lösung eindeutig festzulegen, aber die Menge der Lösungen bzw. Lösungsalgorithmen erheblich verkleinern können.

- Dazu gehört z.B. die Forderung nach der Invarianz¹ des Klassifikationsverfahrens gegenüber solchen Datentransformationen, die wegen des Skalenniveaus der Ausgangsdaten zulässig sind².
- Unterstellt man den zu klassifizierenden Daten stochastischen Charakter, so ist es naheliegend, Stetigkeit der Ergebnisse bei geringfügigen Veränderungen der Ausgangsdaten, die als Zufallsschwankungen angesehen werden können, zu verlangen³.

1 Vgl. FRIEDMAN, H.P., RUBIN, J.: On some Invariant Criteria for Grouping Data. Journal of the American Statistical Association 62 (1967), S. 1159 - 1178.

2 Diesbezügliche Angaben finden sich in den Tabellen 3.2, 3.3 und 3.5.

3 Vgl. DEGENS, P.O.: Clusteranalyse auf topologisch-maßtheoretischer Grundlage. Dissertation, München 1978, und DEGENS, P.O.: Konsistenzforderung für die automatische Klassifikation. a.a.O.

- Einen ganzen Katalog von 9 "Zulässigkeits"bedingungen für Klassifikationsverfahren stellten FISHER/van NESS¹ auf; darunter sind z.B. die Forderung nach Konvexität der Lösung² und die Forderung, daß sich, wenn man von einer Lösung mit n Klassen eine Klasse wegläßt, dieselbe Lösung (eingeschränkt auf die verbleibenden $n-1$ Klassen) ergeben soll.

Untersucht man jedoch die meistgebrauchten numerischen Klassifikationsverfahren hinsichtlich dieser Bedingungen, dann stellt sich z.B. heraus, daß das Verfahren mit den schlechtesten empirischen Erfahrungen, nämlich Single-Linkage, vergleichsweise recht gute mathematische Eigenschaften aufweist, während die empirisch erfolgreichste Methode, nämlich ein Austauschalgorithmus mit dem Varianzkriterium als Zielfunktion, nicht einmal skaleninvariant ist. Daher konnten sich Bedingungen der o.g. Art als Auswahlkriterium für numerische Klassifikationsverfahren auch kaum gegen das subjektive Ermessen des Benutzers durchsetzen und werden - bis auf gelegentlich die Skaleninvarianz und Konvexität - kaum beachtet.

Für die Bevorzugung eines menschlichen Urteils vor einer theoretischen Bedingung gibt es auch gute Gründe:

- Klassifikation ist eine elementare menschliche Fähigkeit und es ist anzunehmen, daß das intellektuelle Klassifizieren innerhalb seiner Kapazitätsgrenzen sehr "effizient" ist³.

1 s. FISHER, L., van NESS, J.W.: Admissible Clustering Procedures. Biometrika 58 (1971), S. 91-104

2 D.i. die Disjunktheit der konvexen Hüllen der Cluster.

3 s. KLIX, F.: Information und Verhalten. 4. Aufl. Berlin 1978, S. 528.

- Das Ergebnis eines Klassifikationsprozesses soll i.d.R. von Menschen weiterverarbeitet werden und muß deshalb dem menschlichen Begriffsvermögen zugänglich sein. Eine Klasseneinteilung jedoch, die nur noch durch Hyperflächen im $\mathbb{R}^m$, d.h. durch komplizierte Gleichungen, und nicht mehr verbal beschrieben werden kann, ist als Mittel der Datenaufbereitung ungeeignet, weil sie oft für den Anwender keine begreifbare Information mehr bietet.

Die grundlegende Bedeutung und die Notwendigkeit des Klassifizierens als Intelligenzleistung ergibt sich daraus, daß man in seiner "Umgebung in unzähligen verschiedenen Situationen unzählbar viele verschiedene Objekte"¹ wahrnimmt und "unterschiedliche Situationen ... in bestimmtem Umfange eine gleichartige Beantwortung"² erfordern. In Hinsicht darauf, daß es zur Nutzung der einmal gemachten Erfahrungen nötig sein wird, in neuen Situationen die gleichartig zu beantwortenden Situationen wiederzuerkennen, ist die Bildung von Klassen, die eine gleiche Reaktion erfordern, ein notwendiges Mittel, den Überblick zu behalten und damit reaktions- und lernfähig zu bleiben. Bei der Klassifizierung wird gleichzeitig von den für die Beantwortung einer Situation irrelevanten Merkmalen abstrahiert, sodaß im Laufe eines Lernprozesses gleichzeitig durch Ausfilterung verhaltensrelevanter Situationsmerkmale die Klassifizierung besser wird und durch bessere Klassifizierung das Erkennen relevanter Merkmale leichter wird.

1 s. KLIX, F.: Information und Verhalten. 4. Aufl. Berlin 1978, S. 544

2 ebenda, S. 555

Viele der Klassifikationen, auf die sich das menschliche Denken stützt, werden allerdings nicht vom einzelnen Menschen gebildet, sondern werden nur übernommen: Die Benutzung der Sprache z.B., deren einzelne Begriffe nichts anderes sind als Klassen von Gegenständen, Eigenschaften usw.¹, erlernt ein Kind durch trial und error von der Umgebung, wobei es kaum die Möglichkeit hat, Begriffsneubildungen durchzusetzen. Es liegt also eigentlich kein Begriffsbildungsprozeß sondern ein Begriffserwerbsprozeß vor.

Die Neubildung einer Klasseneinteilung läßt sich nur dort beobachten, wo noch keine Erfahrungen vorliegen, die zur Differenzierung relevanter und irrelevanter Merkmale hätten führen können. Solche Situationen können z.B. auftreten, wenn durch die Entwicklung neuer Beobachtungstechniken plötzlich viele bisher unbekannte einzellige Lebewesen oder eine Menge verschiedener subatomarer Teilchen entdeckt werden, aber auch dann, wenn z.B. eine vorhandene, feingliedrige Begriffsstruktur für Kinder noch zu kompliziert ist und einstweilen durch eigene einfache Begriffe ersetzt wird.

Obwohl auch in einer solchen Lage die konkreten Anlässe unbekannt sind, zu denen ein Ordnungssystem in Anspruch genommen werden kann, ist sicher, daß das vorhandene Wissen zumindest wiedergefunden werden muß (information retrieval!) bevor es für einen konkreten Zweck ausgewertet werden kann. Unter diesem Aspekt muß man annehmen, daß auch eine erste Klassenbildung nicht wahllos, sondern in einer

1 vgl. dazu EVERITT, B.S.: Unresolved Problems in Cluster Analysis a.a.O., S. 169: "each noun in language is a label used to describe a class of things which have striking features in common", und "unter einem Begriff verstehen wir die Klassifizierung einer Objektmenge auf Grund von Merkmalen". Aus KLIX, F.: Über die notwendige und mögliche Ausdehnung von Begriffsbildungsanalysen auf komplexe Klassifizierungsleistungen. In KLIX, F. (Hrsg): Psychologische Beiträge zur Analyse kognitiver Prozesse. Berlin 1976, S. 163.

erfahrungsgemäß zweckmäßigen Art und Weise vorgenommen wird¹. Leider ist über die Vorgänge beim Klassifizieren wie auch beim Retrieval nur sehr wenig bekannt², sodaß es im Moment nicht möglich wäre, die hohen diesbezüglichen Leistungen des Gehirns durch einen Algorithmus nachzuahmen. Bei Untersuchungen zum lernabhängigen Klassifizieren konnten jedoch einige Regelmäßigkeiten festgestellt werden, die vermutlich auch für das lernunabhängige Klassifizieren von Bedeutung sind:

- Es können "kaum stabile Begriffsstrukturen mit mehr als fünf relevanten Merkmalen aufgebaut werden"³.
- Die Merkmale Form, Anzahl, Größe, Farbe und Lage (in absteigender Reihenfolge) beeinflussen die Ähnlichkeitsempfindung von Versuchspersonen unterschiedlich stark⁴.
- Bei der Klassenbildung werden solche Klassen bevorzugt, die (in der Reihenfolge der Bevorzugung) durch ein einziges Merkmal, durch "und"-Verknüpfung mehrerer Merkmale, durch "oder"-Verknüpfung mehrerer Merkmale und durch bedingte Verknüpfung beschrieben werden können⁵.
- Die Begriffsstrukturen im Langzeitgedächtnis sind hierarchisch organisiert⁶.

1 Dies entspricht der Bildung von Hypothesen beim Erwerb von Begriffsstrukturen bei HOFFMANN, J.: Gedächtnisleistungen in Begriffsbildungsprozessen. In KLIX, F. (Hrsg.): Psychologische Beiträge zur Analyse kognitiver Prozesse. Berlin 1976, S. 128-158.

2 s. SPARCK JONES, K., KAY, M.: Linguistik und Informationswissenschaft. München 1976, S. 31 ff.

3 s. KLIX, F.: Über die notwendige und mögliche Ausdehnung von Begriffsbildungsanalysen auf komplexe Klassifizierungsleistungen. a.a.O., S. 165.

4 s. HOFFMANN, J.: Gedächtnisleistungen in Begriffsbildungsprozessen. a.a.O., S. 165.

5 ebenda S. 146.

6 s. KLIX, F.: Strukturelle und funktionelle Komponenten des menschlichen Gedächtnisses. In KLIX, F. (Hrsg.): Psychologische Beiträge zur Analyse kognitiver Prozesse. Berlin 1976, S. 79.

Im folgenden soll nun untersucht werden, ob es auch für das lernunabhängige Klassifizieren Regeln gibt, also Regeln für das erste Klassifizieren unbekannter Objektmengen, wobei noch keinerlei Erfahrungen über Zweckmäßigkeit einer bestimmten Klasseneinteilung dieser Objektmenge zur Verfügung stehen, sondern nur die allgemeinen Erfahrungen über zweckmäßiges Klassifizieren an sich.

5.2 EXPERIMENTE ZUR KLASSIFIKATION

In den Experimenten sollen jeweils von einer Reihe von Versuchspersonen Gegenstände in Klassen eingeteilt werden, um durch den Vergleich der verschiedenen Ergebnisse festzustellen, in welchen Beziehungen die Ergebnisse übereinstimmen und in welchen nicht.

Wenn zu solchen Experimenten Gegenstände benutzt würden, die den Versuchspersonen bereits vorher bekannt sind, also beispielsweise eine Menge verschiedener Kraftfahrzeuge, was als Untersuchungsgegenstand eines Marktforschers wohl realistisch wäre, dann würden sich daraus drei wesentliche Gefahren ergeben:

- Die meisten uns umgebenden Dinge werden durch eine so große Zahl von Merkmalen beschrieben, daß die Rolle eines einzelnen Merkmals aus dem Ergebnis nicht herausisoliert werden könnte. Zudem sind viele Merkmale, wie z.B. die Form einer Karosserie gar nicht meßbar.
- Bei bekannten Dingen bringen die Versuchspersonen zwangsläufig Vorkenntnisse und Meinungen - z.B. die Erinnerung an den schlechten Service in der Vertragswerkstatt einer bestimmten Automarke - in die Klassifikation mit ein, die nicht erkennbare und den meßbaren Merkmalen nicht zurechenbare Wirkungen hervorrufen können.
- Bei vertrauten Dingen besteht die Gefahr, daß die Versuchspersonen keine eigene Klassifikation erstellen, sondern lediglich eine bereits bekannte Klassenstruktur reproduzieren, die sich möglicherweise schon in entsprechenden Begriffen wie z.B. Mittelklassewagen, Sportwagen, Luxuslimousine niedergeschlagen hat.

Um also sicherzustellen, daß wirklich ein Klassifikationsprozeß beobachtet werden kann und nicht die Repro-

duktion einer bereits bestehenden Struktur, und um kontrollierte Versuchsbedingungen herzustellen, müssen Objekte mit wenigen meßbaren Merkmalen gewählt werden, denen von den Versuchspersonen keine weitere Bedeutung zugemessen wird. Dazu wurden zunächst wieder Zufallszahlen erzeugt, die dann zur Konstruktion einfacher geometrischer Figuren wie Rechtecke, Dreiecke usw. benutzt wurden, indem die Maße dieser Figuren den Zufallszahlen entsprechend gewählt wurden¹.

In der ersten Stufe der Experimente wurden den Probanden Objekte mit lediglich einem veränderlichen Merkmal vorgelegt:

25 Rechtecke mit der konstanten Breite 1 cm und Längen von x_i cm, $1 \leq i \leq 25$, die einer Normalverteilung $N_1(11;25)$ insofern genügen, als durch die Werte der Verteilungsfunktion $F_N(x_i/11;25)$ das Intervall $[0,02; 0,98]$ in 24 gleiche Teile geteilt wird; d.h. die x_i wurden so bestimmt, daß

$$F_N(x_i/11;25) = 0,02 + (i-1)0,04 \quad \text{für } 1 \leq i \leq 25 \quad (5.1)$$

gilt. Die in Abb. 5.1 abgebildeten Rechtecke wurden 17 Personen in Form von Papierstreifen ungeordnet vorgelegt mit der Aufforderung, sie so in Gruppen einzuteilen, daß die Rechtecke in einer Gruppe möglichst ähnlich sind. Die Zahl der zu bildenden Gruppen wurde nicht vorgegeben.

1 Solche Figuren werden auch als Metroglyphen bezeichnet und ebenso wie z.B. ANDREW's Sine Curves, das sind Kurven mit variablen Parametern, oder die immer häufiger benutzten CHERNOFF Faces, das sind stilisierte Gesichter, deren Proportionen von den Daten festgelegt werden, zur Repräsentation multivariater Daten verwendet. vgl. BRUCKNER, L.A.: On Chernoff Faces. In: WANG, P.C.C. (Hrsg.): Graphical Representation of Multivariate Data. New York, San Francisco, London 1978, S. 93 - 122.

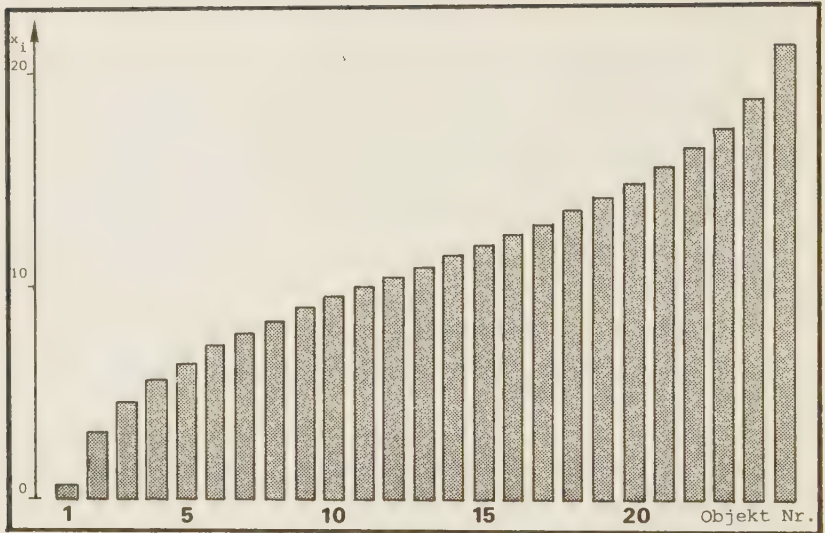


Abb. 5.1: Versuchsobjekte im Versuch 1

Dabei wie auch bei allen späteren Versuchen konnte zunächst die Beobachtung gemacht werden, daß trotz der Unschärfe der Aufgabenstellung, die im Wort "ähnlich" liegt, die Aufgabe fast immer sofort verstanden und meist ohne längeres Überlegen in Angriff genommen wurde. In nur sehr wenigen Fällen wurden die Streifen vor dem Klassifizieren der Länge nach sortiert. Meist wurden zunächst diejenigen Streifen zusammengelegt, die sich nur sehr wenig unterscheiden, und diejenigen mit extremen Längen herausgelegt. Die übrigen wurden dann nach und nach zugeordnet.

Wie das Histogramm in Abb. 5.2 zeigt, wurde keine Gruppenzahl bevorzugt, sondern es wurden gleichmäßig oft 4 bis 8 Gruppen und nur zweimal mehr als 8 Gruppen gebildet.

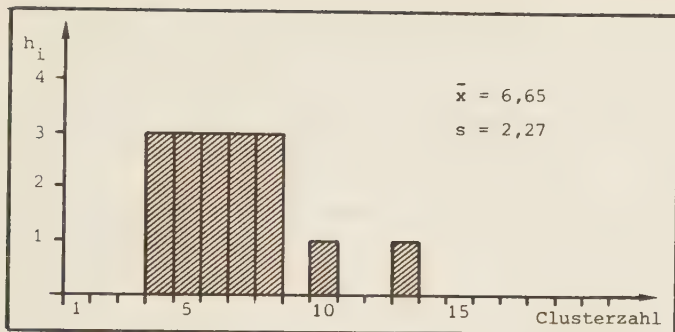


Abb. 5.2: Häufigkeitsverteilung der Gruppennzahlen im Versuch 1

Die Mächtigkeit der einzelnen Gruppen wurde einem χ^2 -Test¹ auf Gleichheit unterworfen; danach mußte beim Irrtumsniveau $\alpha=0,5$ (!) die Nullhypothese der Gleichmächtigkeit noch in 9 Fällen, bei $\alpha=0,01$ jedoch nur noch in 2 Fällen abgelehnt werden. Die Ablehnungen waren meist darauf zurückzuführen, daß die betreffenden Probanden die am Rande liegenden Objekte (s. Abb. 5.1), die sich stärker unterscheiden als die in der Mitte liegenden, in kleinere Gruppen unterteilt haben, was eigentlich nicht anders zu erwarten war.

Größere Längenunterschiede werden also offensichtlich auch wahrgenommen und führen zu einer feineren Klasseneinteilung. Benutzt man jedoch die Häufigkeit h_{ij} , mit der die Streifen i und j , $1 \leq i, j \leq 25$, in dieselbe Gruppe eingeordnet wurden, als Maß für die von den Probanden wahrgenommene Ähnlichkeit der Streifen i und j und dementsprechend $d_{ij} = 17 - h_{ij}$ als Maß für den wahrgenommenen Unterschied,

¹ s. BLEYMÜLLER, J., GEHLERT, G., GÜLICHER, H.: Statistik für Wirtschaftswissenschaftler. a.a.O., S. 123 ff.

dann zeigt sich, daß der wahrgenommene Unterschied nicht zu der tatsächlichen Differenz der Längen $|x_i - x_j|$ proportional ist. Trägt man nämlich in einem Diagramm die tatsächlichen Längen x_i gegen die "wahrgenommenen Längen" y_i ab, die aus den wahrgenommenen Unterschieden der benachbarten Streifen j und $j+1$ als

$$y_i = \sum_{j < i} d_{jj+1} \quad \text{für } 2 \leq i \leq 25, \quad (5.2)$$

$$y_1 = 0$$

bestimmt werden können, so ergibt sich (s. Abb. 5.3), daß die wahrgenommene Länge einen s-förmigen Verlauf annimmt.

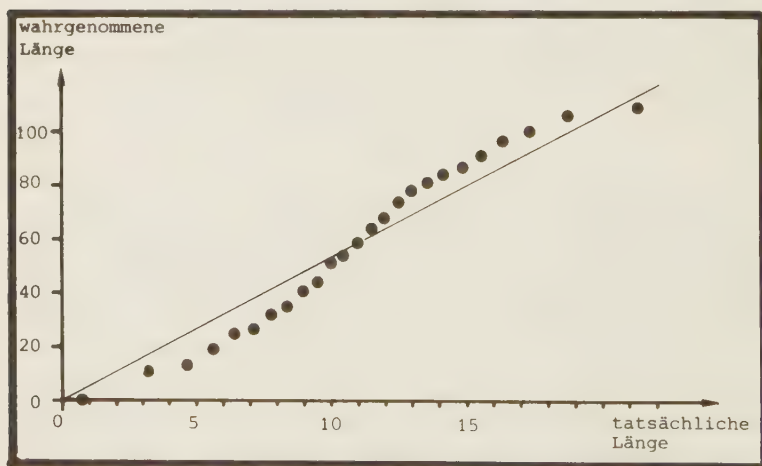


Abb. 5.3: Tatsächliche und wahrgenommene Länge in Versuch 1

Offenbar werden die Längenunterschiede zu beiden Seiten eines Mittelwertes immer schwächer wahrgenommen. Transformiert man die tatsächlichen Längen x_i dagegen durch

$$x_i \longrightarrow \ln\left(\frac{|x_i - \bar{x}|}{s} + 1\right) \operatorname{sign}(x - \bar{x}) \quad , \quad (5.3)$$

dann liegen die Punkte (x_i, y_i) fast genau auf einer Geraden (Abb. 5.4).

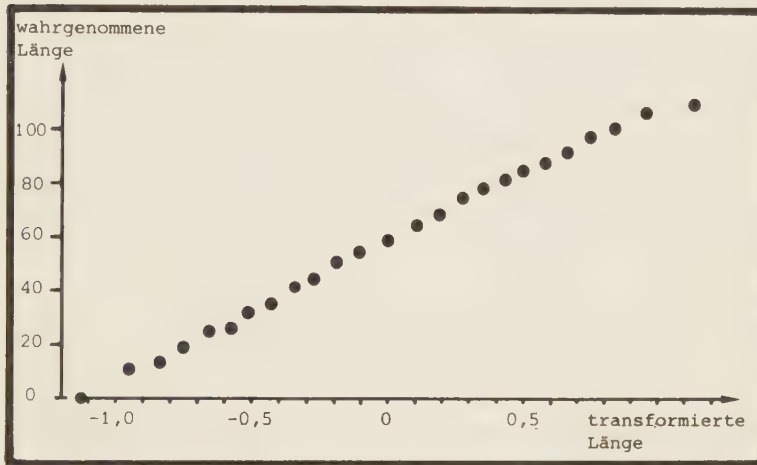


Abb. 5.4: Transformierte und wahrgenommene Länge in Versuch 1

Darin kommt zweierlei zum Ausdruck:

- Bei der Wahrnehmung wird offenbar ein Mittelwert der Reizstärken gebildet und nicht die absolute Ausprägung x_i , sondern die Abweichung vom Mittelwert gemessen. Diese "Nullpunktbildung" ist in der Psychophysik bereits längere Zeit bekannt und spielt nicht nur bei optischen sondern "bei allen eindimensionalen sensorischen Urteilen"¹ eine Rolle.
- Die größeren Merkmalsunterschiede werden nicht im gleichen Maße wie die kleineren berücksichtigt, sondern

1 s. KLIX, F.: Information und Verhalten, a.a.O., S. 245 ff.

durch eine logarithmische Transformation abgeschwächt.

Auch dieser Vorgang ist in Form des WEBER-FECHNERSchen-Gesetzes¹ in der Psychophysik bereits bekannt.

Damit wird deutlich, daß die bekannten Gesetzmäßigkeiten der Wahrnehmung auch für die Klassifikation von Bedeutung sind.

Die Ergebnisse bei einer vorgegebenen Objektmenge aus 50 Streifen der Breite 5 cm, deren Länge als Zufallszahlen aus einer Normalverteilung mit $\mu = 10$ und $\sigma^2 = 25$ erzeugt wurden, waren praktisch die gleichen; bemerkenswert ist daran, daß sich trotz der Verdoppelung der Objektzahl die Zahl der gebildeten Gruppen nur unwesentlich erhöht hat (s. das Histogramm in Abb. 5.5).

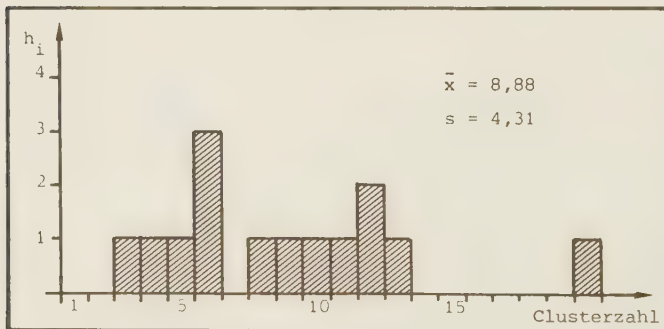


Abb. 5.5: Häufigkeitsverteilung der Gruppenzahl in Versuch 2

In einer zweiten Experimentstufe wurden Objekte mit zwei veränderlichen Merkmalen verwendet: 25 rechtwinklige Dreiecke, bei denen jeweils ein Innenwinkel α_i als Zufallszahl aus einer $N_1(45;225)$ -Verteilung und die Länge x_i

1 s. z.B. SIXTL, F.: Meßmethoden der Psychologie. Weinheim/Bergstraße 1967, S. 96 ff.

der anliegenden Kathete als Zufallszahl aus einer $N_1(10; 25)$ -Verteilung bestimmt wurden. Die Punkte (x_i, α_i) bilden das in Abb. 5.6 wiedergegebene Streuungsdiagramm.

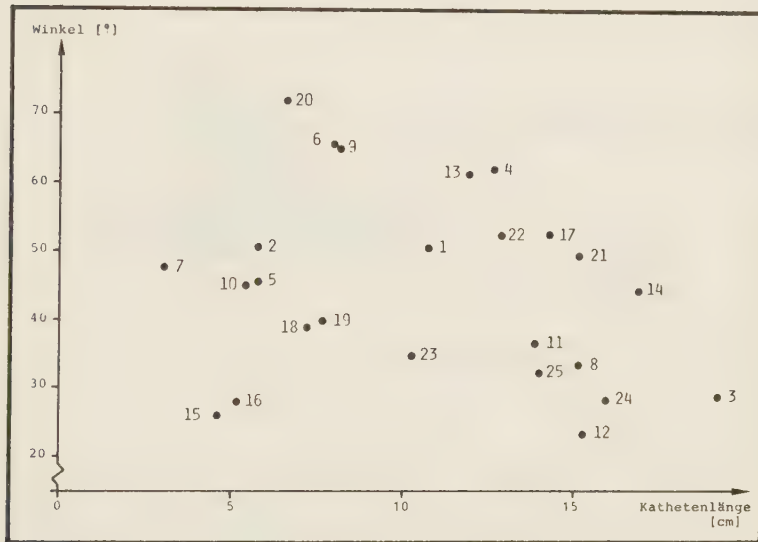


Abb. 5.6: Streuungsdiagramm von Winkeln und Kathetenlängen in Versuch 3

Um das bereits bei diesen einfachen Objekten offenkundig mühselige Sortieren bei ungeordneter Vorlage der Dreiecke, das zeitraubend ist und durch Sortierfehler (z.B. falsch gebildete Reihenfolgen der Kathetenlängen) die Klassifikation beeinflusst, zu vermeiden, wurden die Dreiecke vorsortiert und in der zweidimensionalen Anordnung in Abb. 5.7 präsentiert. Durch einen vorangehenden Test, bei dem einige Personen dieselbe Objektmenge in unterschiedlichen Anordnungen vorgelegt wurde, konnte die Befürchtung, daß die vorgegebene Anordnung Einfluß auf die Klassenbildung haben könnte, weitgehend ausgeräumt werden. Zudem sind die

Dreiecke in Abb. 5.7 nicht den Koordinaten (x_i, α_i) entsprechend wie in Abb. 5.6 angeordnet, wodurch ein Ähnlichkeitsmaß suggeriert werden könnte, sondern in ein Raster mit gleichen Abständen so eingeordnet, daß lediglich die ordinale Reihenfolge der x_i und α_i in der Anordnung eingehalten wird.

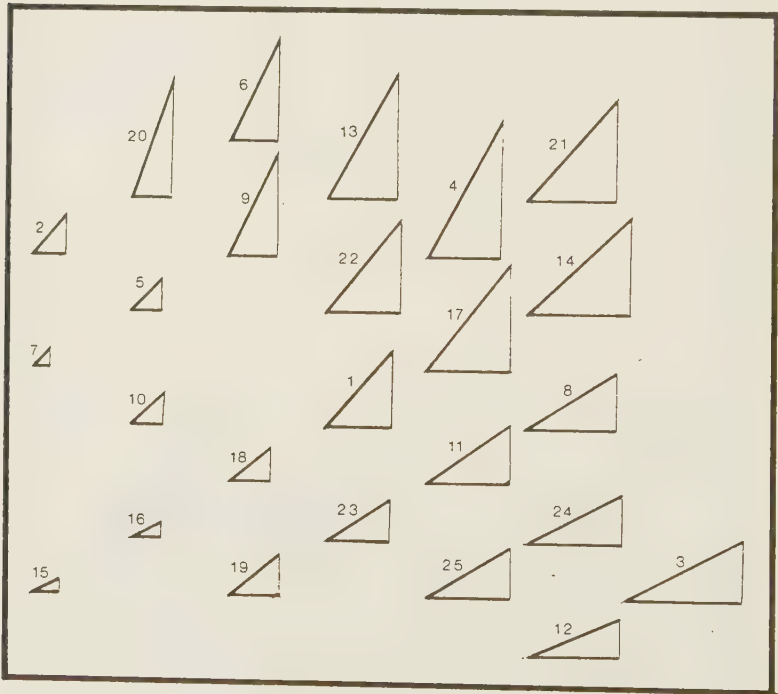


Abb. 5.7: Anordnung der Dreiecke im Versuch 3

Auch hier wurde von 17 Personen zunächst ohne Angabe einer Gruppenzahl klassifiziert. Die gewählten Gruppenzahlen bewegten sich, obwohl zwei Variable mehr Differenzierungsmöglichkeiten bieten als eine, im gleichen

Rahmen wie beim ersten Versuch (vgl. das Histogramm in Abb. 5.8). Dabei waren hier die Mächtigkeiten der Gruppen

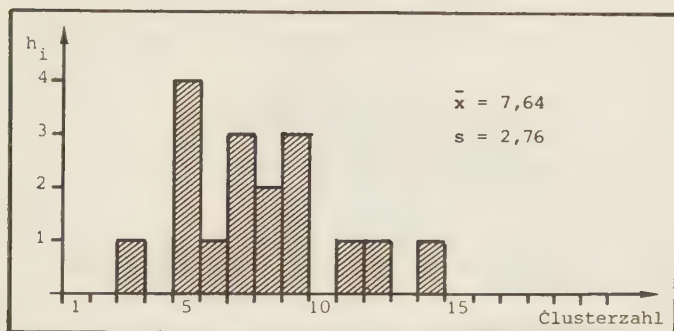


Abb. 5.8: Häufigkeitsverteilung der Gruppenzahlen in Versuch 3

noch gleichmäßiger. Im χ^2 -Test mußte die Hypothese der Gleichmächtigkeit aller Gruppen beim Irrtumsniveau $\alpha=0,5$ nur in 5 von 17 Fällen und beim Irrtumsniveau $\alpha=0,01$ nur noch einmal abgelehnt werden, wobei teilweise extrem kleine Werte der Prüfgröße χ^2 auftraten. Die Darstellung der Lösungen im Merkmalsraum zeigte außerdem, daß die konvexen Hüllen der gebildeten Gruppen bis auf wenige Ausnahmen disjunkt waren.

Die augenfälligste Beobachtung war jedoch die starke Dominanz der Form, d.h. hier des Winkels, über die Größe, in diesem Fall also die Kathetenlänge der Dreiecke. Bildet man die Häufigkeiten h_{ij} , mit der die Dreiecke i und j in dieselbe Gruppe eingeordnet wurden, im Streuungsdiagramm (Abb. 5.6) durch entsprechend starke Verbindungs-
linien ab, so ergibt sich die Abb. 5.9.

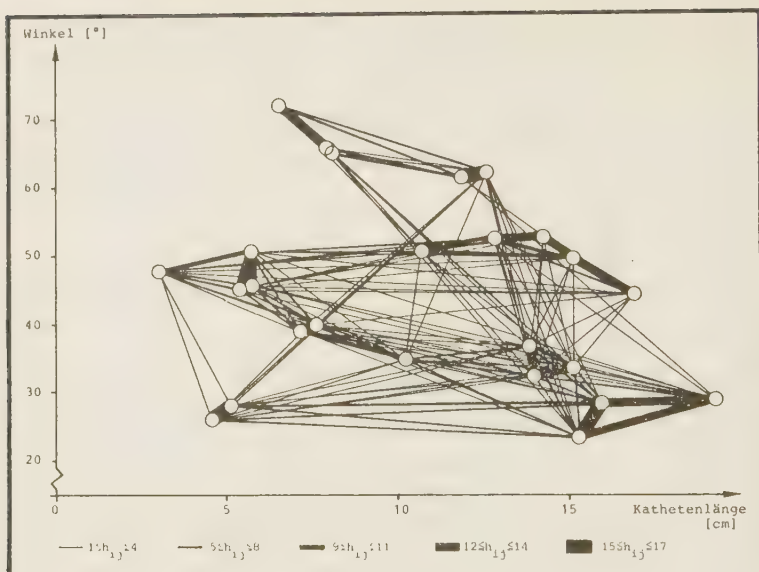


Abb. 5.9: Häufigkeiten h_{ij} , mit der die Dreiecke i und j in die gleiche Gruppe i,j eingeordnet wurden

Daran wird erkennbar, daß einige Dreiecke von fast allen Personen in dieselbe Gruppe eingeordnet wurden, während andere von keiner Person derselben Gruppe zugeordnet wurden. Dabei werden offenbar die Größenunterschiede wesentlich weniger berücksichtigt (erkennbar an den zahlreichen und starken horizontalen Verbindungen in Abb. 5.9) als die Formunterschiede (schwache vertikale Verbindungen).

Dies kann man jedoch noch besser an einem zweiten Versuch mit der gleichen Objektmenge sehen (Versuch 4), bei dem 21 Personen jeweils 3 Gruppen bildeten. Übereinstimmend wurden von allen Personen die Objekte 5,7,10 bzw. 6,9,20 bzw. 15,16 bzw. 3,18,11,12,23,24,25 jeweils in dieselbe Gruppe eingeordnet (s. Abb. 5.10). Über die Zusammengehörigkeit dieser Elemente herrschte

also keinerlei Zweifel, obwohl die übrigen Elemente unterschiedlich zugeordnet wurden.

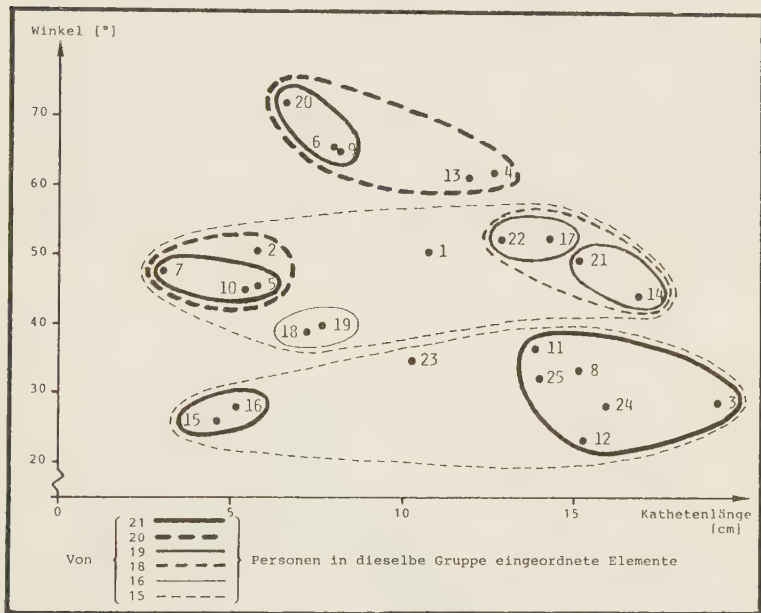


Abb. 5.10: Übereinstimmende Gruppierungen von mehreren Versuchspersonen

Verbindet man weiterhin auch alle Elemente miteinander, die nur von 20 (19, 18 usw.) der 21 Personen übereinstimmend derselben Gruppe zugeordnet wurden, so entstehen zusammenhängende Teilmengen, die in Abb. 5.10 durch Umrandung gekennzeichnet worden sind. Die Gestalt dieser Gebilde macht wiederum sichtbar, daß Dreiecke mit unterschiedlichen Kathetenlängen häufiger zusammen eingeordnet wurden als solche mit unterschiedlichen Winkeln. Offenbar lösen Formunterschiede hier erheblich größere Unähnlichkeitsempfindungen als Größenunterschiede aus und werden daher beim Klassifizieren als trennendes Merkmal bevorzugt.

Auch bei der Betrachtung einzelner Partitionen wie in Abb. 5.11 entsteht der Eindruck, daß die beiden Merkmale beim Klassifizieren nicht zugleich verwendet werden, sondern daß bei der Bildung weniger Klassen nur das dominante Merkmal benutzt wird, und erst bei größeren Klassenzahlen das zweite Merkmal zur weiteren Aufteilung hinzugezogen wird. Dies entspricht der bereits aus der Psychologie bekannten Tatsache, daß zur Begriffsbildung einzelne Merkmale bzw. einfache Verknüpfungen (im Sinne der BOOLEschen Algebra) wie "und" (also hier z.B. $\alpha \leq 40^\circ \wedge x \geq 10 \text{ cm}$) komplizierteren Abgrenzungsmöglichkeiten vorgezogen werden¹.

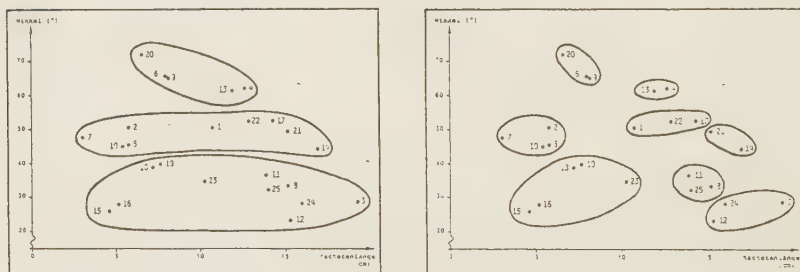


Abb. 5.11: Klassifizierungen derselben Person mit 3 und 8 Klassen

Auch in den Experimenten mit drei veränderlichen Merkmalen wurde das Angebot an Merkmalen und damit an Differenzierungsmöglichkeiten regelmäßig durch Auswahl von Merkmalen oder Transformationen verringert. Die im Dreiecksexperiment beobachtete Dominanz eines Merkmals sollte bei diesem Experiment nach Möglichkeit vermieden werden, indem drei stochastisch unabhängige Variable mit derselben $N_1(3;1)$ -Verteilung durch drei gleichwertige Merkmale dargestellt wurden. Dazu wurden je ein schwarzer, blauer

¹ vgl. S. 128

und roter Papierstreifen, deren Längen Zufallszahlen aus der genannten Verteilung entsprachen, zu einem dreiarmigen Stern nach dem Muster der Abb. 5.12 zusammengefügt und auf ein quadratisches Kärtchen so aufgeklebt, daß die Mittelpunkte des Sterns und des Kärtchens zusammenfielen. 21 Personen wurden nun aufgefordert, die 25 Sterne zu klassifizieren, wobei in 12 Fällen die Klassenzahl freigestellt und in 9 Fällen auf 6 festgelegt wurde.

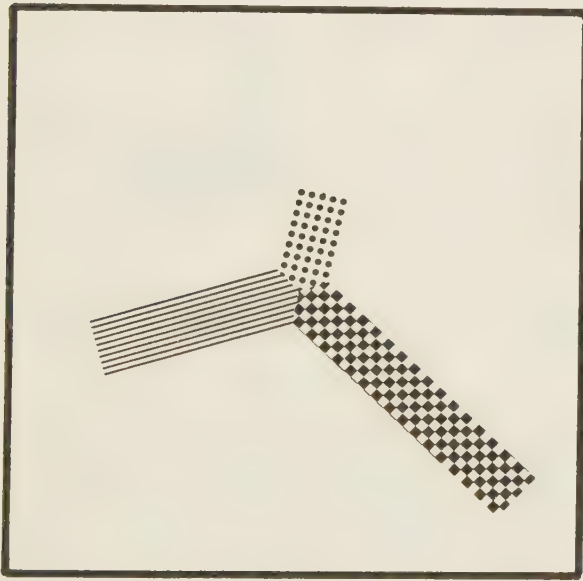


Abb. 5.12: Versuchsobjekt im Versuch 5. Die Farben sind durch Muster  (schwarz),  (rot) und  (blau) ersetzt.

Die Menge der 25 Sterne wirkte auf die Versuchspersonen nach eigenen Angaben zunächst verwirrend, sodaß es ihnen schwerfiel, sofort eine Klasseneinteilung zu entwerfen. In der Regel konnten sie erst nach einigen Minuten des

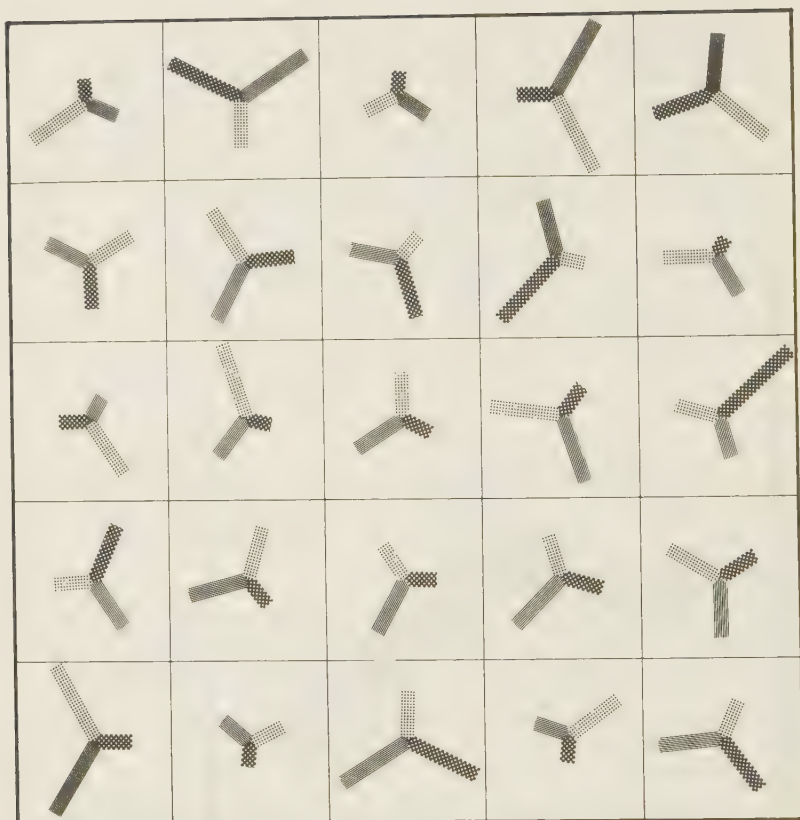


Abb. 5.13: Versuchsobjekte im Versuch 5;  $\hat{=}$ schwarz,  $\hat{=}$ rot,  $\hat{=}$ blau.

Probierens ein Konzept finden, das ihnen aber in vielen Fällen auch dann noch unbefriedigend erschien, weil ihnen selbst klar war, daß es ihnen nicht gelungen war, alle drei Merkmale zu berücksichtigen.

Die frei wählbaren Klassenzahlen lagen wiederum in demselben Bereich wie zuvor (s. Abb. 5.14) und der χ^2 -Test auf Gleichmächtigkeit der Klassen führte bei einem Irrtumsniveau von $\alpha=0,5$ in 8 von insgesamt 21 Fällen und bei $\alpha=0,1$ nur noch in 2 Fällen zur Ablehnung der Nullhypothese.

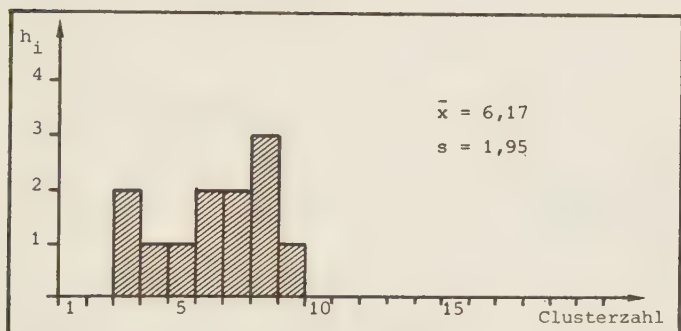


Abb. 5.14: Häufigkeitsverteilung der Gruppenzahlen in Versuch 5

Dies ist umso erstaunlicher, als im Gegensatz zum vorigen Experiment hier von mehreren Probanden völlig unterschiedliche Konzepte der Klassenbildung entwickelt wurden. Um diese erkennen zu können, wurden die gefundenen Lösungen in einem Modell des dreidimensionalen Merkmalraumes (s. Abb. 5.15) dargestellt, in welchem jeder Stern durch eine Kugel repräsentiert wird, deren Raumkoordinaten den Merkmalsausprägungen des Sterns entsprechen. Dadurch wurde es bei einem Teil der Ergebnisse ermöglicht, die Klassifikationsregeln zu erkennen.

Aus Abb. 5.15(b) wird deutlich, daß in diesem wie in einigen anderen Fällen überhaupt nur die Arme des Sterns mit einer bestimmten Farbe (hier die roten) zur Einteilung herangezogen wurden und die anderen völlig unbe-

rücksichtigt blieben. Im Gegensatz zu der personenunabhängigen Bevorzugung der Form vor der Größe bei den Dreiecken wurden hier alle drei Farben von verschiedenen Personen als alleiniges Klassifizierungsmerkmal verwendet.

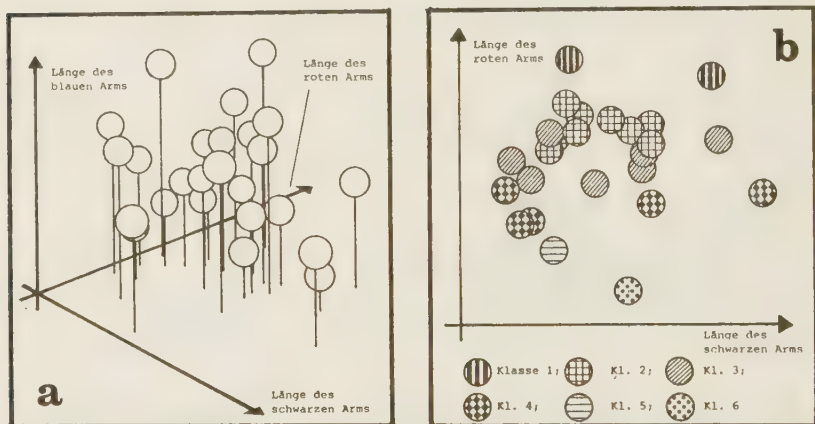


Abb. 5.15: Modell des Merkmalsraumes von der Seite (a) und von oben (b) mit markierter Klasseneinteilung in (b)

Gegenüber dem klaren Konzept der Klassenbildung in Abb. 5.15(b) erschienen einige andere Lösungen im Modell zunächst völlig chaotisch. Erst die Befragung der Versuchspersonen nach den Klassifizierungsregeln erklärte die scheinbare Unordnung: Nachdem sie festgestellt hatten, daß sie die drei Farben nicht gleichgewichtig und zugleich verarbeiten konnten, hatten sie beschlossen, von der Farbe völlig abzusehen und nur noch die Form zu beachten. Dies führte zu den Klassen mit 3, 2, 1 oder 0 "langen Armen", wobei eventuell die Unterscheidung von "langen", "mittleren" und "kurzen" Armen noch zu einer Fein-

teilung herangezogen wurde. Durch die Nichtbeachtung der Farbe wurde also der ursprünglich 3-dimensionale Merkmalsraum auch hier in einen 1-dimensionalen Merkmalsraum transformiert, um die Klassifikationsaufgabe dadurch zu vereinfachen.

Diejenigen Lösungen, die unter Einbeziehung aller drei Farben entstanden waren, setzten sich meist aus Gruppen zusammen, die im Modell mit einem Ende auf den Ursprung des Koordinatensystems wiesen und mit wachsender Entfernung vom Ursprung fächerartig breiter wurden (s.Abb. 5.16).

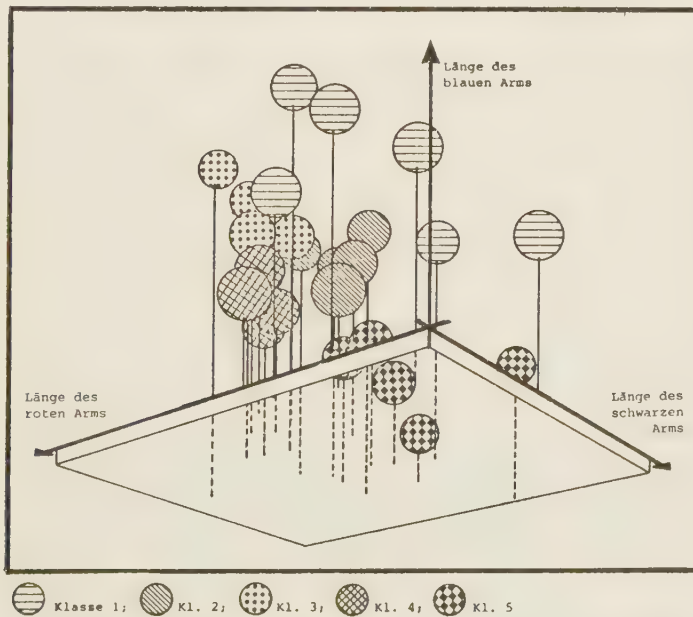


Abb. 5.16: Modell des Merkmalsraumes aus der Sicht vom Ursprung des Koordinatensystems mit markierter Klasseneinteilung

Die Probanden gaben in diesen Fällen an, die Längenverhältnisse der Arme mit verschiedenen Farben verglichen zu haben. Das Ergebnis waren Klassen, die sich recht gut durch algebraische Verknüpfungen von ">", "<" und "≅" - Aussagen beschreiben ließen. Die Beschreibung der Klassifikation in Abb. 5.16 hätte dann etwa folgende Form (mit s = Länge des schwarzen Arms, r (rot) und b (blau) analog):

1. Klasse: $r < b$
 2. Klasse: $r \equiv b \equiv s$
 3. Klasse: $r \equiv b \wedge r > s$
 4. Klasse: $r > b \wedge r > s$
 5. Klasse: $r > b \wedge r \leq s$
- (5.4)

Die BOOLEsche Algebra war bereits bei den Begriffsbildungsprozessen als bevorzugte Methode zur Verknüpfung von Merkmalen in Erscheinung getreten und bestätigt hier nur noch einmal den Zusammenhang zwischen Begriffsbildung und Klassifikation. Bemerkenswert an der Klasseneinteilung (5.4) ist jedoch darüberhinaus, daß nicht jede Merkmalskala für sich in Abschnitte unterteilt wird, was insgesamt einer Zerlegung des Merkmalsraumes in Quader entspräche; stattdessen werden die drei Merkmale sofort zueinander ins Verhältnis gesetzt. D.h. es werden eigentlich nicht die drei Längen r, b und s betrachtet, sondern die Quotienten b/r, b/s und r/s. Da von diesen drei Quotienten jeweils der dritte aus den beiden anderen bestimmt werden kann, genügt es, nur zwei davon zu betrachten. Der dreidimensionale Merkmalsraum kann also durch die Abbildung

$$(r, b, s) \mapsto \left(\frac{b}{r}, \frac{r}{s} \right) \quad (5.5)$$

in einen zweidimensionalen Raum abgebildet werden, wobei - bis auf einen konstanten Faktor - keine Information verlorengeht.

Gleichzeitig wird jedoch, wie aus (5.4) ersichtlich ist, die unendlich große Zahl der möglichen Ausprägungen von r , b und s auf nur 3 zu unterscheidende Ausprägungen der Verhältnisse b/r und r/s verringert: " <1 ", " >1 " und " ≈ 1 ". Die Transformation (5.5) wird also von einer Verringerung des Skalenniveaus begleitet. Dadurch bekommen die im dreidimensionalen Modell relativ schwer zu beschreibenden Gruppen eine recht einfache Gestalt (s. Abb. 5.17).

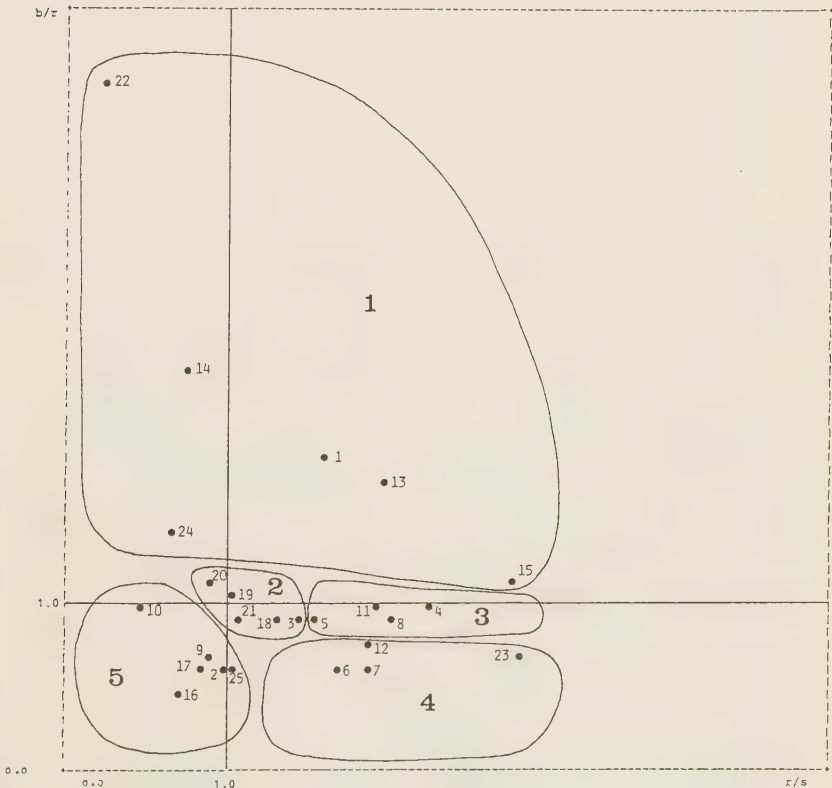


Abb. 5.17: Klasseneinteilung aus Abb. 5.16 im transformierten Merkmalsraum

Im Grunde liegt also auch hier eine Vereinfachung des Merkmalsraumes durch Verringerung der Dimensionszahl vor. Im Gegensatz zu den zuvor beschriebenen Lösungen wurden bei den hier beschriebenen Klassifizierungen aber nicht einfach Informationen weggelassen, sondern die Vereinfachung wurde durch geeignete Verarbeitung aller Merkmale erreicht.

Die gleichen Beobachtungen konnten auch an einer anderen Menge von Sternen gemacht werden, die aber so konstruiert wurden, daß sie eine im Modell gut erkennbare Gruppenstruktur aufwiesen (s. Abb. 5.18): Eine kleine Gruppe aus 6 Sternen mit kurzen roten und langen schwarzen Armen (in Abb. 5.18 ) , eine Gruppe aus 9 Sternen, die sich von kurzen roten und schwarzen mit gleichzeitig langen blauen Armen diagonal durch den Merkmalsraum bis hin zu

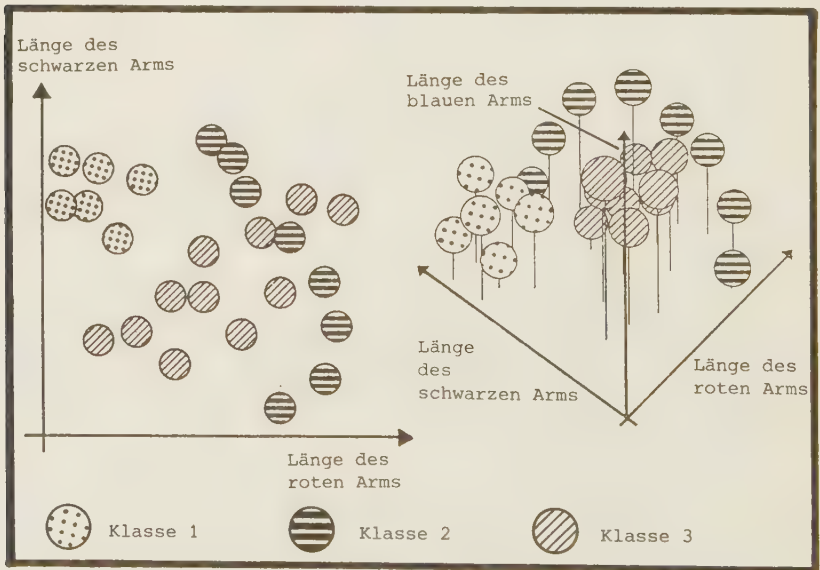


Abb. 5.18: Modell des Merkmalsraumes in Versuch 6 aus zwei verschiedenen Perspektiven

langen roten und schwarzen und kurzen blauen Armen erstreckt und die Gestalt eines Ellipsoids hat () , sowie eine Kette aus 8 Sternen, die bogenförmig durch den Merkmalsraum verläuft (). Mit dieser Objektmenge sollte einerseits untersucht werden, ob die Gruppenstruktur die Klassifizierung beeinflusst. Andererseits wurde die gleiche Objektmenge mit den im dritten Kapitel vorgestellten numerischen Verfahren geclustert, um die Ergebnisse numerischer und intellektueller Klassifikation vergleichen zu können.

Ein Maß zum Vergleich der Klassifikationsergebnisse mit der vorgegebenen richtigen Gruppierung wurde folgendermaßen konstruiert:

Aus der richtigen Partition $\mathcal{A}$ kann eine Ähnlichkeitsmatrix $\mathbf{S}$ hergeleitet werden, indem

$$s_{ij} = \begin{cases} 1 & \text{falls Objekt } i \text{ und } j \text{ Elemente} \\ & \text{derselben Klasse sind} \\ 0 & \text{falls Objekt } i \text{ und } j \text{ Elemente} \\ & \text{verschiedener Klassen sind} \end{cases} \quad (5.6)$$

für $1 \leq i, j \leq n$ gesetzt wird. Genauso erhält man aus der mit $\mathcal{A}$ zu vergleichenden Partition $\mathcal{A}^*$ eine Ähnlichkeitsmatrix $\mathbf{S}^*$. Da die s_{ij} nur zwei mögliche Ausprägungen besitzen, kann nur eines der in Abschnitt 3.2¹ genannten Ähnlichkeitsmaße für zweiwertige nominalskalierte Merkmale zur Berechnung eines Ähnlichkeitswertes s zwischen $\mathbf{S}$ und $\mathbf{S}^*$ und damit auch zwischen $\mathcal{A}$ und $\mathcal{A}^*$ herangezogen werden. Dieses Vorgehen hat insbesondere den Vorteil, daß auch Partitionen mit unterschiedlicher Länge, also mit unterschiedlichen Klassenzahlen, auf diese Weise verglichen werden können. Hier wurde als Ähnlichkeitsmaß der

1 s.S. 63 ff.

Koeffizient von TANIMOTO benutzt:

$$s = \frac{a}{a + (b+c)} \quad (5.7)$$

mit a = Anzahl der Objektpaare, die in beiden Partitionen Element derselben Klasse sind

und $b+c$ = Anzahl der Objektpaare, die in einer Partition derselben Klasse, in der anderen Partition verschiedenen Klassen angehören.

Die Klassifikationsergebnisse der 15 Personen in diesem Versuch wurden genau wie die Ergebnisse des vorigen Versuchs von den verschiedenen Methoden geprägt, den Merkmalsraum zu vereinfachen. Die Struktur der Objektmenge, die nur durch gleichzeitiges Betrachten aller Merkmale erkennbar werden kann, wurde dadurch völlig zurückgedrängt und hatte so gut wie keinen Einfluß auf die Ergebnisse. Die Werte des TANIMOTO-Koeffizienten lagen dementsprechend zwischen 0,135 und 0,343 (der beste erreichbare Wert ist 1,0).

Dagegen lagen die jeweils besten Ergebnisse, die mit Zielfunktionen im Austauschalgorithmus bei Vorgabe von 2 bis 10 Clustern erzielt wurden, zwischen 0,406 und 0,522 und die jeweils besten Ergebnisse der hierarchischen Verfahren zwischen 0,476 und 0,891. Die numerischen Verfahren liefern demzufolge völlig andere Ergebnisse als die Personen und schneiden in Hinsicht auf die Strukturerkennung sämtlich besser ab, wobei die hierarchischen Verfahren deutlich vorn liegen. Diese Rangfolge der Ergebnisse wird auch durch die Verwendung eines anderen Ähnlichkeitsmaßes (wie z.B. des Simple-Matching Koeffizienten oder des ϕ -Koeffizienten¹) nicht wesentlich beeinflusst.

1 s. S. 70

Verfahren		bester Wert von s	zugehörige Klassenzahl
Hierarchische Verf.	Single Linkage	0,891	4
	Complete Linkage	0,476	2
	Average Linkage	0,484	2
	Weighted Av. Link.	0,484	2
	Median Verf.	0,577	3
	Centroid Verf.	0,543	3
	WARD's Verf.	0,639	2
	Flexible Strategie	0,490	4
Zielfunktionen	Varianzkriterium	0,406	2
	Abstandssummenk.	0,412	6
	WILKS' Lambda	0,466	2
	Größter Eigenwert	0,466	2
	HOTELLING- T_0^2	0,559	4
	Adaptive Distanzen	0,522	3
Versuchspersonen		0,343	5

Tab. 5.1: Größte Werte des TANIMOTO-Koeffizienten beim Vergleich der richtigen Partition mit den künstlichen Partitionen

5.3 ZUSAMMENFASSUNG

Aus dem Zusammenhang der Experimente des letzten Abschnitts ergeben sich die folgenden Beobachtungen:

Unabhängig von der Anzahl der klassifizierten Objekte, der Anzahl der Variablen, dem Vorhandensein oder Nichtvorhandensein einer Datenstruktur und vor allem unabhängig von der personenspezifischen Technik der Klassenbildung ist in allen Experimenten eine deutliche Tendenz festzustellen, etwa gleich mächtige Klassen zu bilden.

In allen Experimenten mit 25 Objekten werden Partitionen mit 3 bis 14 Klassen gebildet, wobei der Mittelwert bei 6,98 und die Standardabweichung bei 2,33 lag. Innerhalb dieses Bereichs wurde keine Klassenzahl besonders bevorzugt. Lediglich bei einem Versuch mit 50 Objekten wurden bis zu 20 Klassen gebildet, wobei sich der Mittelwert auf 8,88 und die Standardabweichung auf 4,31 erhöhten. Die Zahl der Klassen scheint also von der Zahl der Merkmale nur unwesentlich beeinflusst zu werden und steigt auch mit größeren Objektzahlen nur unterproportional an.

Die aus der Psychologie bekannten Gesetzmäßigkeiten der Wahrnehmung und Verarbeitung von Reizen wie die Mittelwertbildung, logarithmische Transformation von Reizunterschieden und die Dominanz der Form über die Größe eines Gegenstandes bei der Beurteilung von Ähnlichkeiten sind an der Gestaltung der Klassen erkennbar beteiligt.

Höhere Ansprüche an die Klassifikationsleistung, die durch mehrere veränderliche Merkmale entstehen, werden umgangen, indem die Dimensionszahl des Merkmalsraumes durch vereinfachende Transformationen reduziert wird. Dabei wird meistens die nicht mehr zu bewältigende Informationsmenge einfach außer acht gelassen; nur in wenigen Fällen konnten noch 3 Variable gemeinsam verarbeitet werden.

Viele Personen empfanden schon die Klassifizierung nach zwei Merkmalen als schwierig.

Die Partitionen waren bis auf wenige Ausnahmen konvex. Bei den Versuchen mit mehreren Merkmalen bezieht sich dies gegebenenfalls auf den transformierten, also vereinfachten Merkmalsraum, in welchem letztlich klassifiziert wurde.

Die Partitionen, die unter Berücksichtigung mehrerer Merkmale gebildet worden waren, können durch die Verknüpfung von Bedingungen nach den einfachen Regeln der BOOLEschen Algebra beschrieben werden, die auch zur Bildung sprachlicher Begriffe bevorzugt werden.

Offenbar gibt es keine allgemein bevorzugte Strategie beim Klassifizieren, aber einige Regeln, die unabhängig von der individuellen Strategie aus verschiedenen Gründen eingehalten werden. An die Gesetzmäßigkeiten der Wahrnehmung ist natürlich jeder gebunden; aber es ist kein zwingender Grund bekannt, z.B. gleich große Gruppen oder Gruppenzahlen in bestimmten Größenordnungen zu bevorzugen. Dies könnte damit zusammenhängen, daß beim Suchen nach einer bestimmten Information in einem Speicher, dessen Inhalt in Klassen gegliedert ist, die Zahl der Entscheidungen, die notwendig sind, um die richtige Klasse und die spezielle Information zu finden, von der Größe und Zahl der Klassen abhängig ist. Bei größeren Datenmengen wird dann zwangsläufig eine hierarchische Organisationsstruktur notwendig, um weder die Zahl noch den Umfang der Klassen zu groß werden zu lassen. Tatsächlich gibt es Hinweise¹ darauf, daß das Gedächtnis, das ja

1 s. z.B. LOFTUS, E.F.: Semantic Memory Retrieval: Some Data and a Model. In: Formal Aspects of Cognitive Processes, Lecture Notes in Computer Science Bd. 22, Berlin, Heidelberg, New York 1975, S. 41-54, oder KLIX, F.: Strukturelle und funktionelle Komponenten des menschlichen Gedächtnisses. a.a.O., S. 77 ff.

trotz seines immensen Fassungsvermögens ein sehr effizientes Retrieval ermöglicht, hierarchisch organisiert ist, was bedeuten würde, daß das Prinzip der hierarchischen Klassifikation zumindest für Zwecke des Information Retrieval unter allgemeinsten Voraussetzungen als wichtiger Gesichtspunkt für die Gestaltung eines Klassifikationssystems anzusehen wäre.

Die Bestrebungen, einen mehrdimensionalen Merkmalsraum in einen möglichst einfachen Raum zu transformieren, sind wohl einfach die Folgen biologischer Leistungsgrenzen. Da diese individuell verschieden sind, ist auch der Zwang zur Vereinfachung verschieden stark. Die Beschränkung auf wenige Merkmale macht es aber auf jeden Fall erforderlich, die Merkmale in eine Reihenfolge ihrer Wichtigkeit zu bringen, um jeweils die wichtigsten auswählen zu können. Diese Reihenfolge kann sich aber zunächst an nichts anderem orientieren, als an den Erfahrungen über die Relevanz bestimmter Merkmale in der Vergangenheit, die sich z.B. in der regelmäßigen Bevorzugung der Form vor der Größe äußern. Solche Erfahrungen sollten aber auch dann berücksichtigt werden, wenn Beschränkungen hinsichtlich der Zahl der zu verarbeitenden Merkmale nicht bestehen.

Die Fähigkeit, mehrdimensionale Datenstrukturen zu erkennen, scheint bei Menschen nicht sehr ausgeprägt zu sein, was hauptsächlich darauf zurückgeführt werden muß, daß in den meisten Fällen die Mehrdimensionalität der Merkmalsräume gar nicht verarbeitet wird. Obwohl auch die numerischen Methoden nur spezifische Datenstrukturen erkennen können, schneiden sie in dieser Beziehung immer noch deutlich besser ab als die Versuchspersonen. Dabei muß man allerdings in Betracht ziehen, daß die Ergebnisse beim intellektuellen Klassifizieren im Gegensatz zur numerischen Klassifikation von der Art der Repräsentation

der Daten abhängen und daß das Klassifizieren von Gegenständen etwas anderes ist als das Klassifizieren der Punkte im Merkmalsraum, die diese Gegenstände repräsentieren. Dieselbe Gruppenstruktur, die von den Versuchspersonen z.B. anhand der Sterne nicht erkannt wurde, konnte nämlich anhand des räumlichen Modells meist schnell identifiziert werden. Das durch die geometrische Nähe der Punkte im Merkmalsraum suggerierte Ähnlichkeitsmaß stimmt also offensichtlich nicht mit den Ähnlichkeitsmaßen überein, welche die Personen für die Gegenstände selbst gefunden hatten. Dies ist noch einmal ein deutlicher Hinweis darauf, daß durch den Abstraktionsschritt

Gegenstand $\longrightarrow$ Vektor von Merkmalsausprägungen

wesentliche Informationen verlorengehen können, sodaß die aus den Vektoren gewonnenen Ergebnisse für den Gegenstandsbereich bedeutungslos sind.

6 SCHLUSSBEMERKUNG

"Clusteranalyse" ist eine Sammelbezeichnung für Methoden, größere Objektmengen in homogene Teilmengen zu zergliedern. Verschiedenartige Vorstellungen über das, was die "Homogenität" einer solchen Teilmenge, eines "Clusters", ausmacht, haben zu einer Fülle unterschiedlicher Modellansätze und Techniken geführt, sodaß der Anwender heute einer nicht mehr zu übersehenden Anzahl verschiedener Verfahren gegenübersteht, mit denen die unterschiedlichsten Ergebnisse produziert werden können.

Ziel dieser Arbeit war es, festzustellen, ob und mit welchen Kriterien die Nützlichkeit der verschiedenen Methoden bzw. die Güte der Ergebnisse beurteilt werden kann, um so die Auswahl objektiver gestalten zu können. Dazu erwies es sich als notwendig, zunächst die Aufgabe der Clusteranalyse zu präzisieren, damit die Ergebnisse daran gemessen werden können, inwiefern die gestellte Aufgabe gelöst worden ist. Da die Aufgabe zum einen in der Analyse einer den Daten innewohnenden Struktur, zum anderen im Erstellen einer nach irgendeinem Kriterium zweckmäßigen Klasseneinteilung bestehen kann, muß die Beurteilung auch je nach Art der Anforderung unter diesen zwei unterschiedlichen Aspekten getroffen werden.

Experimente zur Strukturanalyse zeigten, daß beteiligte Personen kaum in der Lage waren, eine Gruppenstruktur in einem dreidimensionalen Merkmalsraum zu erkennen. Die Fähigkeit, ein Clusteranalyseergebnis in höherdimensionalen Merkmalsräumen zu beurteilen, muß ihnen daher weitgehend abgesprochen werden. Experimente mit numerischen Verfahren wiederum zeigten, daß der vielschichtige Begriff "Struktur" für einen einzelnen Algorithmus zu komplex ist, weil jeder Algorithmus konstruktionsbedingt nur ein bestimmtes Kennzeichen einer Struktur erkennen kann. Da aber in der Regel vorher nicht bekannt ist, durch welches Kennzeichen sich die Struktur des vorliegenden Datensatzes auszeichnet, und dem An-

wender ein Urteil über die Richtigkeit der verschiedenen möglichen Ergebnisse nicht abverlangt werden kann (s.o.), kann eine Entscheidung darüber, ob eine Clusterlösung eine tatsächlich vorhandene Struktur wiedergibt oder ein methodisches Artefakt darstellt, nur anhand der erreichten Werte der Zielfunktionen bzw. Heterogenitätsmaße getroffen werden, die innerhalb der Algorithmen schon zur Steuerung des Zuordnungsprozesses benutzt werden. Entsprechende Entscheidungsverfahren sind jedoch zur Zeit noch kaum entwickelt und erwiesen sich in den Experimenten größtenteils als sehr unzuverlässig.

Die wichtigsten Ansatzpunkte für eine weitere Entwicklung der Clusteranalyse als Mittel zur Aufdeckung von Datenstrukturen liegen daher in der Präzisierung des zu lösenden Problems, d.h. insbesondere in einer exakten Definition des Begriffs "Struktur", und in der Bereitstellung von geeigneten Kriterien, die es ermöglichen zu entscheiden, welche Clusteranalysemethode bzw. welche Lösung bei einem bestimmten Datensatz adäquat ist. Dies umfaßt natürlich auch die Möglichkeit, daß in der Folge neuer Modellansätze - z.B. aus der Graphentheorie - auch weiterhin neue Methoden entwickelt werden. Die Verbesserung der bekannten Methoden wie z.B. die Beschleunigung der Algorithmen oder die weitere Annäherung an das globale Optimum einer Zielfunktion kann demgegenüber solange nur von untergeordneter Bedeutung sein, wie nicht grundsätzlich geklärt werden kann, inwiefern die angewandte Methode überhaupt eine im Sinne der verfolgten Absichten vernünftige Lösung liefern kann.

Für die Klassifizierung, bei der nicht die Analyse einer Datenstruktur im Vordergrund steht, gibt es nahezu keine Bewertungsgesichtspunkte, sofern nicht ein Gütekriterium wie z.B. eine Kostenfunktion o.ä. aus dem unmittelbaren Situationzusammenhang abgeleitet werden kann. Obwohl es auch in den Fällen, in denen der Zweck einer Klassifizierung nicht

in die Form einer Zielfunktion gebracht werden kann, mehr oder weniger sinnvolle Einteilungen gibt, ist kaum etwas darüber bekannt, welche Eigenschaften von Klasseneinteilungen und welche Klassifizierungsstrategien sich erwartungsgemäß am häufigsten als nützlich erweisen.

Die Experimente, bei denen Personen ohne Angabe eines Verwendungszweckes aufgefordert wurden, ihnen unbekannte Objektmengen zu klassifizieren, zeigten, daß zwar verschiedene Personen nach verschiedenen Gesichtspunkten klassifizieren, daß aber einige Richtlinien wie z.B. das Bestreben nach gleichen Klassengrößen mehrheitlich eingehalten wurde. Von diesen wird anscheinend intuitiv erwartet, daß sie unabhängig von dem späteren Verwendungszweck vorteilhaft sind, sodaß man annehmen kann, daß solche Regeln auch als Bestandteil eines numerischen Klassifikationsalgorithmus zu dessen Effizienz beitragen würden. Durch die genauere Erforschung menschlicher Klassifizierungstechniken könnten daher sicher weitere Hinweise für die Gestaltung von Algorithmen gewonnen werden. Trotz der stellenweisen Überlegenheit von Computern, was z.B. Speicherkapazität und Verarbeitungsgeschwindigkeit angeht, müssen nämlich Gedächtnisleistungen wegen der hohen Flexibilität beim Speichern, Verarbeiten und Wiederabrufen selbst großer Informationsmengen immer noch als vorbildlich und nachahmenswert angesehen werden.

Eine andere Möglichkeit, Hinweise auf notwendige Eigenschaften eines guten Klassifikationssystems zu erhalten, wäre die Untersuchung der vielen bereits bestehenden Systeme, die teilweise durch sehr lange Zeiträume entwickelt wurden und sich bereits bewährt haben: Die Systematik von Pflanzen und Tieren in der Biologie, die Einteilung von Teilchen und Stoffen in Physik und Chemie, Information Retrieval Systeme, Verfahren des Pattern Recognition, etc. etc..

Von denjenigen Eigenschaften dieser Systeme, die sich in allen oder zumindest den meisten Systemen wiederfinden, muß man annehmen, daß sie vom speziellen Gegenstandsbereich unabhängig sind und auch bei anderen Objektmengen von Vorteil sein werden. Deshalb sollten auch die Ergebnisse von numerischen Methoden nach Möglichkeit dieselben Eigenschaften aufweisen.

Auf der Basis einer präziseren Aufgabenstellung bei der Analyse von Datenstrukturen bzw. mit einem Katalog von Eigenschaften, die ein Klassifikationsverfahren erfüllen müßte, könnte schließlich die Fülle der vorhandenen Clusteranalysemethoden so geordnet werden, daß dann dem Anwender möglicherweise wirkungsvollere Kriterien an die Hand gegeben werden können, um für sein spezielles Problem ein geeignetes Verfahren auszuwählen.

LITERATURVERZEICHNIS

- ABRAMOWITZ, M.,
STEGUN, I.A.: Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. National Bureau of Standards, Applied Mathematics Series Bd. 55, 10. korrr. Aufl., Washington 1972.
- AHRENS, H.,
LÄUTER, J.: Mehrdimensionale Varianzanalyse, Hypothesenprüfung, Dimensionserniedrigung, Diskrimination. Berlin 1974.
- ANDERBERG, M.R.: Cluster Analysis for Applications. New York, London 1973.
- BERGONIER, H.,
MORIN, B.: Typology at Random. ESOMAR Seminar on Segmentation and Typology, Brüssel 1972.
- BLEYMÜLLER, J.,
GEHLERT, G.: Statistische Formeln und Tabellen. München 1979.
- BLEYMÜLLER, J.,
GEHLERT, G.,
GÜLICHER, H.: Statistik für Wirtschaftswissenschaftler. München 1979.
- BOCK, H.H.: Automatische Klassifikation. Göttingen 1974.
- DERSELBE: Clusteranalyse mit unscharfen Partitionen. In: BOCK, H.H. (Hrsg.): Klassifikation und Erkenntnis III. Frankfurt 1979, S. 137 - 163.
- BRUCKNER, L.A.: On Chernoff Faces. In: WANG, P.C.C. (Hrsg.): Graphical Representation of Multivariate Data. New York, San Francisco, London 1978, S. 93 - 121.
- BRYANT, P.: On Testing for Clusters Using the Sample Covariance. Journal of Multivariate Analysis 5 (1975), S. 96 - 105.
- BÜTTNER, J.: Zur Clusteranalyse. Biometrische Zeitschrift 17 (1975), S. 163 - 179.
- CLEMENS, U.: Mehrdimensionale Skalierung bei gemischten Daten. Dipl.-Arbeit, Münster 1980.
- DEGENS, P.O.: Clusteranalyse auf topologisch-maßtheoretischer Grundlage. Dissertation, München 1978.

- DEGENS, P.O.: Konsistenzforderung für die automatische Klassifikation. In: BOCK, H.H. (Hrsg.): Klassifikation und Erkenntnis III. Frankfurt 1979, S. 31 - 47.
- DURAN, B.S.,
ODELL, P.L.: Cluster Analysis, A Survey. Lecture Notes in Economics and Mathematical Systems, Vol. 100, Berlin, Heidelberg, New York 1974.
- ENGELMAN, L.,
HARTIGAN, J.A.: Percentage Points of a Test for Clusters. Journal of the American Statistical Association 64 (1969), S. 1647 - 1648.
- EVERITT, B.S.: Unresolved Problems in Cluster Analysis. Biometrics 35 (1979), S. 169 - 181.
- FISHER, L.,
VAN NESS, J.W.: Admissible Clustering Procedures. Biometrika 58 (1971), S. 91 - 104.
- FISHER, R.A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7 (1936-37), S. 179 - 188.
- FRIEDMAN, H.P.,
RUBIN, J.: On some Invariant Criteria for Grouping Data. Journal of the American Statistical Association 62 (1967), S. 1159 - 1178.
- GOWER, J.C.: A General Coefficient of Similarity and some of its Properties. Biometrics 27 (1971), S. 857 - 874.
- HERDEN, W.,
STEINHAUSEN, D.: Ein Verfahren zur Berechnung eines globalen Minimums beim Varianzkriterium in der Cluster-Analyse. In: SPÄTH, H. (Hrsg.): Ausgewählte Operations Research Software in FORTRAN. München, Wien 1979, S. 115 - 142.
- HOFFMANN, J.: Gedächtnisleistungen in Begriffsbildungsprozessen. In: KLIX, F. (Hrsg.): Psychologische Beiträge zur Analyse kognitiver Prozesse. Berlin 1976, S. 128 - 158.
- HOTELLING, A.: A Generalized T-test and a Measure of Multivariate Dispersion. In: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1959, S. 23 - 41.
- KLIX, F.: Information und Verhalten. 4. Aufl., Berlin 1978.
- DERSELBE: Strukturelle und funktionelle Komponenten des menschlichen Gedächtnisses. In: KLIX, F. (Hrsg.): Psychologische Beiträge zur Analyse kognitiver Prozesse. Berlin 1976, S. 57 - 98.

- KLIX, F.: Über die notwendige und mögliche Ausdehnung von Begriffsbildungsanalysen auf komplexe Klassifizierungsleistungen. In: KLIX, F. (Hrsg.): Psychologische Beiträge zur Analyse kognitiver Prozesse. Berlin 1976, S. 159 - 170.
- KRES, H.: Statistische Tafeln zur multivariaten Analysis. Berlin, Heidelberg, New York 1975.
- KREYSZIG, E.: Statistische Methoden und ihre Anwendungen. 6. Aufl., Göttingen 1977.
- KRUSKAL, J.: The Relationship between Multidimensional Scaling and Clustering. In: VAN RYZIN (Hrsg.): Classification and Clustering. New York, San Francisco, London 1977, S. 17 - 44.
- KUIPER, K.F.,
FISHER, L.: A Monte Carlo Comparison of Six Clustering Procedures. Biometrics 31 (1975), S. 777 - 783.
- LABOVITZ, S.: The Assignment of Numbers to Rank Order Categories. American Sociological Review 35 (1970), S. 515 - 524.
- LANCE, G.N.,
WILLIAMS, W.T.: A General Theory of Classificatory Sorting Strategies I. Hierarchical Systems. The Computer Journal 9 (1966), S. 373 - 380.
- LEHMANN, E.L.: Nonparametrics, Statistical Methods Based on Ranks. San Francisco 1975.
- DERSELBE: Testing Statistical Hypotheses. New York 1959.
- LILLIEFORS, H.W.: On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association 62 (1967), S. 399 - 402.
- LOFTUS, E.F.: Semantic Memory Retrieval: Some Data and a Model. In: Formal Aspects of Cognitive Processes. Lecture Notes in Computer Science Bd. 22, Berlin, Heidelberg, New York 1975, S. 41 - 54.
- MORRISON, D.F.: Multivariate Statistical Methods. New York u.a. 1967

- NABER, A.: Multivariate Analyse von Produktmärkten. Dissertation, Münster 1979.
- PARKS, J.M.: Classification of Mixed Mode Data by R-Mode Factor Analysis and Q-Mode Cluster Analysis on Distance Function. In: COLE, A.J. (Hrsg.): Numerical Taxonomy. London, New York 1969, S. 216 - 219.
- ROHLF, J.F.: Generalization of the Gap Test for the Detection of Multivariate Outliers. Biometrics 31 (1975), S. 93 - 101.
- ROY, S.N.: Some Aspects of Multivariate Analysis. New York 1957.
- SCHALLER, K.: Eine Methodenbank zur Clusteranalyse. In: BOCK, H.H. (Hrsg.): Klassifikation und Erkenntnis III. Frankfurt 1979, S. 79 - 94.
- SCHMITZ, N.,
LEHMANN, F.: Monte-Carlo-Methoden I. Erzeugen und Testen von Zufallszahlen. Meisenheim am Glan 1976.
- SITTERBERG, G.: Multivariate Analyse der Struktur und Entwicklung von Städten. Beiträge zum Siedlungs- und Wohnungswesen und zur Raumplanung Bd. 41, Münster 1977.
- SIXTL, F.: Meßmethoden der Psychologie. Weinheim/Bergstraße 1967.
- SNEATH, P.H.A.,
SOKAL, R.R.: Numerical Taxonomy. San Francisco 1973.
- SODEUR, W.: Empirische Verfahren zur Klassifikation. Stuttgart 1974.
- SPÄTH, H.: Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion. München, Wien 1975.
- DERSELBE: Cluster-Analyse: Konzepte, Rezepte und Anwendungen. Manuskript zu einem Vortrag beim 11. Steiermärkischen Mathematischen Symposium in Graz, 1.-4.10.1979.
- SPÄTH, H.,
MÜLLER, R.: Das Austauschverfahren für die skalierungsinvariante Methode der adaptiven Distanzen in der Cluster-Analyse. In: SPÄTH, H. (Hrsg.): Ausgewählte Operations Research Software in FORTRAN. München, Wien 1979, S. 143 - 163.

- WILKS, S.S.: Certain Generalizations in the Analysis of Variance. Biometrika 24 (1932), S. 471 - 494.
- WITTING, H., Angewandte Mathematische Statistik. Stuttgart 1970.
- ZURMÜHL, R.: Matrizen und ihre Technischen Anwendungen. Berlin, Göttingen, Heidelberg, 4. neubearbeitete Aufl. 1964.

LEBENS LAUF

Siegfried Bergs	Im Kleekamp 6, 4700 Hamm
Geburtstag	20. Mai 1951
Geburtsort	Bottrop
Staatsangehörigkeit	deutsch
Eltern	Else Bergs, geb. Kortmann, und Siegfried Bergs, Steiger
Schulbesuch	1957-1961 Volksschule in Hamm/Marl 1961-1969 Neusprachliches Gymnasium in Marl
Wehrdienst	1969-1971
Studium	1971-1977 Mathematik an der Univer- sität Münster 1977 Examen als Diplom-Mathematiker seit 1978 Betriebswirtschaftslehre an der Universität Münster
Berufsweg	1974-1977 Studentische Hilfskraft am Mathematischen Institut der Univer- sität Münster 1977-1978 Wissenschaftliche Hilfskraft am Institut für Ökonometrie und Wirt- schaftsstatistik der Universität Münster seit 1978 dort Wissenschaftlicher Assistent

